HANDBOOK OF GEOMETRIC TOPOLOGY

Edited by
R.J. Daverman
R.B. Sher

NORTH-HOLLAND

HANDBOOK OF GEOMETRIC TOPOLOGY

HANDBOOK OF GEOMETRIC TOPOLOGY

Edited by

R.J. DAVERMAN

University of Tennessee, Knoxville, Tennessee, USA

R.B. SHER

497 Cedar Ridge Road, Union Hall, Virginia, USA

2002

ELSEVIER

Amsterdam • London • New York • Oxford • Paris • Shannon • Tokyo

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

ISBN: 0-444-82432-4
First edition 2002

Library of Congress Cataloging-in-Publication Data
Handbook of geometric topology/edited by R.J. Daverman, R.B. Sher. – 1st ed.
p. cm.
Includes indexes.
ISBN 0-444-82432-4 (alk. paper)
1. Topology. I. Daverman, Robert J. II. Sher, R. B.
QA161 .H36 2001
514–dc21

2001051281

British Library Cataloguing in Publication Data
Handbook of geometric topology
1. Topology
I. Daverman, R. J. II. Sher, R. B.
516

ISBN 0444824324

⊗ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in The Netherlands

Preface

Geometric Topology focuses on matters arising in special spaces such as manifolds, simplicial complexes, and absolute neighborhood retracts. Fundamental issues driving the subject involve the search for topological characterizations of the more important objects and for topological classification within key classes. Undoubtedly the most famous question of them all is the still unsettled Poincaré Conjecture, dating from 1904, which posits that any simply-connected compact 3-manifold (without boundary) is topologically the 3-sphere. This is a prototypical problem for the subject: within a given class (3-manifolds), do elementary topological properties (simple-connectedness and compactness) yield a strong global conclusion (being the 3-sphere)?

The development of this relatively young subject has been stunning. In the first half of the century the bulk of the attention fell on 3-manifolds, polyhedra and other low-dimensional objects of a seemingly "concrete" nature rooted in our intuitive notions of "space". The 1960s and 1970s saw long strides taken in the analysis of high-dimensional manifolds, including Smale's proof of the h-cobordism theorem and, with it, the solution of the generalized Poincaré Conjecture, topological characterizations of infinite-dimensional manifolds, and classifications of infinite-dimensional manifolds modeled on the Hilbert cube by simple-homotopy type. In the last portion of the 20th century came such results as: the analysis of 4-manifolds, powerfully stoked by Donaldson's gauge-theoretic methods and Freedman's topological analysis of topological handle cancellation; the adoptation of geometric methods (often embodied in the study of manifolds whose universal coverings are familiar geometric objects, but for which the covering transformations are isometries) spurred in dimension 3 by Thurston and carried out in dimensions greater than 4 by Farrell and Jones, among others; a variety of results on 3-manifolds and classical knot-theory emerging from new invariants such as the Jones polynomial; and the emergence of an algebraic-geometric-topological hybrid known as geometric group theory.

This Handbook is intended for readers with some knowledge of Geometric Topology (or even only certain limited aspects of the subject) and with an interest in learning more. It was put together in the hope and belief that graduate students in particular would find it useful. Among other features, it offers perspectives on matters closely studied in times past, such as PL topology, infinite-dimensional topology, and group actions on manifolds, and it presents several chapters on matters of intense interest at the time it was assembled, near the beginning of a new millenium, such as geometric group theory and 3-manifolds (knot theory included) and their invariants. It includes current treatments of vital topics such as cohomological dimension theory, fixed point theory, homology manifolds, invariants of high-dimensional manifolds, mapping class groups, structures on manifolds and topological dynamics. Unfortunately the editors were not able to obtain appropriate coverage of recent important developments in the theory of 4-manifolds.

The editors are grateful for all the help provided them in putting together this volume, especially by the staff at Elsevier Science and by all of the authors who provided chapters for inclusion here.

R.J. Daverman and R.B. Sher

List of Contributors

Adem, A., *University of Wisconsin–Madison, Madison, WI* (Ch. 1)
Bestvina, M., *University of Utah, Salt Lake City, UT* (Ch. 2)
Bonahon, F., *University of Southern California, Los Angeles, CA* (Ch. 3)
Boyer, S., *Université du Québéc à Montréal, Montréal* (Ch. 4)
Bryant, J.L., *Florida State University, Tallahassee, FL* (Ch. 5)
Cannon, J.W., *Brigham Young University, Provo, UT* (Ch. 6)
Chigogidze, A., *University of Saskatchewan, Saskatoon* (Ch. 7)
Davis, J.F., *Indiana University, Bloomington, IN* (Ch. 1)
Davis, M.W., *The Ohio State University, Columbus, OH* (Ch. 8)
Dydak, J., *University of Tennessee, Knoxville, TN* (Ch. 9)
Franks, J., *Northwestern University, Evanston, IL* (Ch. 10)
Geoghegan, R., *SUNY at Binghamton, Binghamton, NY* (Ch. 11)
Ivanov, N.V., *Michigan State University, East Lansing, MI* (Ch. 12)
Lee, K.B., *University of Oklahoma, Norman, OK* (Ch. 13)
Lickorish, W.B.R., *University of Cambridge, Cambridge* (Ch. 14)
Lück, W., *Westfälische Wilhelms-Universität Münster, Münster* (Ch. 15)
Plaut, C., *University of Tennessee, Knoxville, TN* (Ch. 16)
Ratcliffe, J.G., *Vanderbilt University, Nashville, TN* (Ch. 17)
Raymond, F., *University of Michigan, Ann Arbor, MI* (Ch. 13)
Scharlemann, M., *University of California, Santa Barbara, CA* (Ch. 18)
Shalen, P.B., *University of Illinois at Chicago, Chicago, IL* (Ch. 19)
Stark, C.W., *National Science Foundation, Arlington, VA* (Ch. 20)
Sullivan, M.C., *Southern Illinois University, Carbondale, IL* (Ch. 10)
Weinberger, S., *University of Chicago, Chicago, IL* (Ch. 21)

Contents

CHAPTER 1

Topics in Transformation Groups*

Alejandro Adem

Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA

James F. Davis

Department of Mathematics, Indiana University, Bloomington, IN 47405, USA

Contents

*Both authors were partially supported by NSF grants.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Preliminaries

1.1. *Preface*

We will not deal here with the historical background of transformation groups. It suffices to say that they occupy a central rôle in mathematics due to their fundamental importance and ubiquitous nature. Rather we will go straight to the basic objects and examples in the subject and from there describe their development in modern mathematics, emphasizing connections to other areas of algebraic and geometric topology. Our goal is to describe some of the fundamental examples and techniques which make transformation groups an important topic, with the expectation that the interested reader will consult the listed references for a deeper understanding. We feel that the area of transformation groups continues to be a testing ground for new techniques in algebraic and geometric topology, as well as a source of accessible problems for mathematical research. We thus list some of the basic conjectures still open in the subject, although the interested researchers will be left to find the accessible problems on their own. Although aspects of the subject can now be regarded as "classical", our knowledge of group actions on arbitrary compact manifolds is far from complete, even in the case of finite groups. Furthermore it should be said that research on actions and topological invariants of infinite discrete groups is a topic of great current interest, involving diverse techniques from group theory, topology and analysis.

Our presentation is organized as follows: in Section 1 we deal with basic notions and examples, with the conviction that examples are the best approach for introducing transformation groups; in Section 2 we describe the cohomological aspects associated to group actions which are most relevant in algebraic topology; finally in Section 3 we discuss the more geometric aspects of this area. Lists of problems are provided in Sections 2 and 3. Finally we would like to make clear that in this text we present a view of transformation groups which reflects our personal interests, omitting such topics as actions of connected Lie groups, and group actions and low-dimensional topology. In no way do we pretend that this is a comprehensive survey of the subject. Points of view on the contents of such a survey will differ; hopefully our list of references will at least point the reader towards other material that may fail to appear in this brief synopsis.

1.2. *Basic definitions*

A *topological group* is a group which is a Hausdorff topological space, with continuous group multiplication and inversion. Any group can be given the structure of a topological group by equipping the group with the discrete topology. We shall concern ourselves mostly with discrete groups.

A *left action of a topological group G on a Hausdorff space X* is a continuous map

$$\begin{aligned} &G \times X \to X, \\ &(g, x) \mapsto gx, \end{aligned}$$

so that $(gh)x = g(hx)$ and $ex = x$ for all $g, h \in G$ and $x \in X$, where $e \in G$ is the identity. One says that X is a *G-space*. A *G-map* (or *equivariant map*) is a map $f: X \to Y$ between G-spaces which commutes with the G-action, that is, $f(gx) = gf(x)$.

A group action defines a homomorphism

$$\begin{aligned} \theta &: G \to \mathrm{Homeo}(X), \\ g &\mapsto (x \mapsto gx), \end{aligned}$$

where $\mathrm{Homeo}(X)$ is the group of homeomorphisms of X; conversely if G is discrete then any such homomorphism defines a group action. An action is *effective* if $\ker\theta = \{e\}$, that is, for every g there is an x so that $gx \neq x$.

Given a point $x \in X$, define the *orbit* $Gx = \{gx \mid g \in G\} \subset X$. The *orbit space* X/G is the set of all orbits, given the quotient topology under the obvious surjection $X \to X/G$, $x \mapsto Gx$. A group action is *transitive* if X consists of a single orbit Gx. A typical example of a transitive G-space is a homogeneous space $X = G/H$.

Given a point $x \in X$, the *isotropy subgroup* is $G_x = \{g \in G \mid gx = x\} < G$. Two points in the same orbit have conjugate isotropy groups

$$G_{gx} = gG_xg^{-1}.$$

A group action is *free* if for every point $x \in X$, the isotropy group is trivial, that is, $gx \neq x$ for all $x \in X$ and all $g \in G - \{e\}$. A typical example of a free action is the action of the fundamental group $\pi_1(X, x_0)$ of a connected CW complex on its universal cover $\widetilde{X}$. The *fixed-point set* of the G-action on X is defined as the subset $X^G = \{x \in X \mid gx = x \; \forall g \in G\}$.

An action of a locally compact Hausdorff group G on a space X is *proper* (also termed *properly discontinuous* when G is discrete) if for every $x, y \in X$, there are neighborhoods U of x and V of y so that $\{g \in G \mid gU \cap V \neq \phi\}$ has compact closure in G. If a discrete group acts freely and properly on X, then $X \to X/G$ is a covering space. Conversely if Y is path-connected and has a universal cover $\widetilde{Y}$ and if H is a normal subgroup of $\pi = \pi_1(Y, y_0)$, then $G = \pi/H$ acts freely and properly via deck transformations on $X = \widetilde{Y}/H$ with orbit space Y.

1.3. *Examples*

The subject of transformation groups is motivated by examples. In this section we give various natural examples of group actions on manifolds arising from representation theory and geometry. In later sections we will discuss classification results, regularity results (i.e., to what extent do arbitrary actions resemble naturally occurring ones), and the construction of exotic actions.

By a representation of a topological group G, we mean a continuous homomorphism from G to an orthogonal group $O(n)$. Since $O(n)$ acts on a wide variety of spaces, such as $\mathbb{R}^n$, D^n, $\mathbb{S}^{n-1}$, $\mathbb{R}P^{n-1}$ and $G_k(\mathbb{R}^n)$, one obtains a multitude of G-actions from a representation. Likewise a complex representation $G \to U(n)$ gives actions on $\mathbb{C}P^{n-1}$, $G_k(\mathbb{C}^n)$, etc. A group action "arising" from a continuous homomorphism $G \to GL_n(\mathbb{R})$ will be called a *linear action*, however, we won't make that precise. We also remark that any smooth action of a compact Lie group G on a smooth manifold M is locally linear: every $x \in M$ has a neighborhood which is G_x-diffeomorphic to a linear G_x-action on $\mathbb{R}^n$.

Here are some examples of linear actions. Let $\mathbb{Z}/k = \langle T \rangle$ be a cyclic group of order k, let $i_1, \ldots, i_n$ be integers relatively prime to k, and let ζ_k be a primitive kth root of unity. Then $\mathbb{Z}/k$ acts on $\mathbb{S}^{2n-1} \subset \mathbb{C}^n$ via

$$T(z_1, \ldots, z_n) = \left(\zeta_k^{i_1} z_1, \ldots, \zeta_k^{i_n} z_n\right).$$

The quotient space $\mathbb{S}^{2n-1}/(\mathbb{Z}/k)$ is the *lens space* $L(k; i_1, \ldots, i_n)$. The quaternion eight group $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ is a subgroup of the multiplicative group of unit quaternions

$$\mathbb{S}^3 = \left\{a + bi + cj + dk \in \mathbb{H} \mid a^2 + b^2 + c^2 + d^2 = 1\right\}$$

and $\mathbb{S}^3/Q_8$ is called the *quaternionic space form*. These are examples of *linear spherical space forms* $\mathbb{S}^{n-1}/G$, which arise from representations $\rho : G \to O(n)$ so that for every $g \in G - \{e\}$, $\rho(g)$ has no $+1$ eigenvalues. The quotient $\mathbb{S}^{n-1}/G$ is then a complete Riemannian manifold with constant sectional curvature equal to $+1$, conversely every such manifold is a linear spherical space form. More generally, complete Riemannian manifolds with constant sectional curvature are called *space forms*, they are quotients of $\mathbb{S}^n$, $\mathbb{R}^n$, or $\mathbb{H}^n$ by a discrete group of isometries acting freely and properly. An excellent discussion is found in Wolf [174].

For a Riemannian manifold M of dimension n, the group of isometries is a Lie group, whose dimension is less than or equal to $n(n+1)/2$; equality is realized only when M is the sphere $\mathbb{S}^n$, real projective space $\mathbb{R}P^n$, Euclidean space $\mathbb{R}^n$, or hyperbolic space $\mathbb{H}^n$. These results are classical, see [103]. If M is compact, so is the isometry group $\operatorname{Isom}(M)$. Conversely if a compact Lie group G acts effectively and smoothly on a compact manifold M, then by averaging one can put a Riemannian metric on M so that G acts by isometries. For a closed, smooth manifold M, the *degree of symmetry of* M is the maximal dimension of a compact Lie group which acts effectively and smoothly on M. A systematic study of the degree of symmetry of exotic spheres is found in [85].

Proper actions of infinite discrete groups have been widely studied, especially proper actions on Euclidean space. For example, a *crystallographic group* Γ is a discrete subgroup of the rigid motions of Euclidean space $\operatorname{Isom}(\mathbb{R}^n)$ so that $\Gamma \setminus \operatorname{Isom}(\mathbb{R}^n)/O(n) = \Gamma \setminus \mathbb{R}^n$ is compact. More generally, a proper action of a discrete group on Euclidean space is determined by a discrete subgroup Γ of a Lie group G, where G has a finite number of components. Then Iwasawa decomposition theory shows that there is a maximal compact subgroup K, unique up to conjugacy, with G/K diffeomorphic to $\mathbb{R}^n$. Given a locally compact group G, subgroups Γ and K with Γ discrete and K compact, then Γ acts properly on the homogeneous space $X = G/K$. Suppose Γ and Γ' are two subgroups of a Lie group G, abstractly isomorphic as groups. The question of rigidity [126] asks if they are conjugate subgroups of G. The Bieberbach rigidity theorem asserts that crystallographic groups are rigid, in the weaker sense that two isomorphic crystallographic groups are conjugate by an affine map of Euclidean space. For many examples of proper actions see [144].

Group actions also play an important part in basic constructions for homotopy theory. Let X denote a topological space with a basepoint: using this point we can obtain natural inclusions $X^n \to X^{n+1}$, where the symmetric groups act by permutation of coordinates so that these maps are equivariant. The n-fold symmetric product on X is defined to be the quotient space $SP^n(X) = X^n/\Sigma_n$, and the infinite symmetric product is defined to be the limit $SP^\infty(X) = \lim_{n\to\infty} SP^n(X)$. A remarkable theorem due to Dold and Thom [64] asserts that $\pi_i(SP^\infty(X)) \cong H_i(X, \mathbb{Z})$. A related construction is the configuration space on n unordered points in X, defined by $C_n(X) = (X^n - D)/\Sigma_n$, where D consists of all n-tuples $(x_1, \dots, x_n)$ such that $x_i = x_j$ for some $i \neq j$ (note that the Σ_n-action is free). These spaces arise in many situations in geometry, topology and physics. In particular if $X = \mathbb{C}$, then $\pi_1(C_n(X)) = B_n$, Artin's braid group on n strings. More sophisticated constructions involving the symmetric groups give rise to models for infinite loop spaces (see [117]).

Covering spaces give natural examples of group actions; we illustrate this with knot theory. If K is a knot (= embedded circle) in $\mathbb{S}^3$, and n is a positive integer, there is a unique epimorphism $\pi_1(\mathbb{S}^3 - K) \to \mathbb{Z}/n$. The corresponding n-fold cyclic cover can be completed to a cyclic branched cover $X_n \to \mathbb{S}^3$, that is, $\mathbb{Z}/n$ acts on a closed 3-manifold X_n so that $(X_n/(\mathbb{Z}/n), X_n^{\mathbb{Z}/n}/(\mathbb{Z}/n))$ is $(\mathbb{S}^3, K)$. The homology group $H_1(X_n)$ was the first systematic knot invariant [9,148,51].

Exotic (yet naturally occurring) examples of group actions are given by symmetries of Brieskorn varieties [25, Part V, §9]. For a non-zero integer d, let $V = V_d^{2n}$ be the complex variety in $\mathbb{C}^{n+1}$ given as the zero set of

$$z_0^d + z_1^2 + \cdots + z_n^2 = 0.$$

The orthogonal group $O(n)$ acts on V fixing the first coordinate and acting on the last n coordinates via matrix multiplication. The variety V has a singularity only at the origin, so

$$\Sigma = \Sigma_d^{2n-1} = V \cap \mathbb{S}^{2n+1}$$

is a smooth $(2n-1)$-dimensional submanifold of $\mathbb{S}^{2n+1}$ and Σ is $O(n)$-invariant. Brieskorn investigated the algebraic topology of Σ and found that when n and d are both odd, Σ_d^{2n-1} is homeomorphic to the sphere, but may have an exotic differential structure. For even n, $H_{n-1}(\Sigma) = \mathbb{Z}/d$, and $H_i(\Sigma) = 0$ for $i \neq 0, n-1, 2n-1$. Also Σ_d^3 is the Lens space $L(d; 1, 1)$. In particular, using the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \in O(3),$$

there is a $\mathbb{Z}/2$-action on Σ_d^5 with fixed set $\Sigma_d^3 = L(d; 1, 1)$. Since there is no exotic differential structure on $\mathbb{S}^5$, this gives a non-linear $\mathbb{Z}/2$-action on $\Sigma_d^5 \cong \mathbb{S}^5$ for odd $d > 2$. One can also construct non-linear actions on $\mathbb{S}^4$. This stands in contrast to lower dimensions. It is not difficult to show that all smooth actions of finite groups on $\mathbb{S}^1$ and $\mathbb{S}^2$ are homeomorphic to linear actions, and this is conjectured for $\mathbb{S}^3$. It has been shown [124] that all smooth actions of a finite cyclic group on $\mathbb{S}^3$ with fixed set a knot are homeomorphic to a linear action; this was conjectured by P.A. Smith.

1.4. *Smooth actions on manifolds*

A *Lie group* is a topological group which is a smooth ($= C^{\infty}$) manifold where group multiplication and inversion are smooth maps. A *smooth action of a Lie group G on a smooth manifold M* is an action so that $G \times M \to M$ is a smooth map. For a discrete group G, there is the corresponding notion of a *PL-action on a PL-manifold.*

The following proposition is clear for discrete G, and requires a bit of elementary differential topology [63, II, 5.2] for the general case.

PROPOSITION 1.1. *For a smooth, proper, free action of a Lie group G on a manifold M, the orbit space M/G admits a smooth structure so that the quotient map $M \to M/G$ is a submersion.*

To make further progress we restrict ourselves to compact Lie groups. To obtain information about M/G we have a theorem of Gleason [73].

THEOREM 1.1. *Suppose a compact Lie group G acts freely on a completely regular space X. Then $X \to X/G$ is a principal G-bundle.*

We will give a nice local description (the Slice theorem) of a smooth action of a compact Lie group. The key results needed are that G-invariant submanifolds have G-tubular neighborhoods and that orbits Gx are G-invariant submanifolds. We will only sketch the theory; for full proofs the reader is referred to Bredon [25] and Kawakubo [100].

THEOREM 1.2. *Suppose a compact Lie group G acts smoothly on M. Any G-invariant submanifold A has a G-invariant tubular neighborhood.*

SKETCH OF PROOF. A G-invariant tubular neighborhood is a smooth G-vector bundle η over A and a smooth G-embedding

$$f : E(\eta) \to M$$

onto a open neighborhood of A in M such that the restriction of f to the zero section is the inclusion of A in M.

We first claim that M admits a Riemannian metric so that G acts by isometries. By using a partition of unity, one can put an inner product $\langle\langle , \rangle\rangle$ on the tangent bundle $T(M)$. To obtain a G-invariant metric, one averages using the Haar measure on G

$$\langle v, w \rangle = \int_G \langle\langle gv, gw \rangle\rangle \, \mathrm{d}g.$$

Then the exponential map

$$\exp : W \to M$$

is defined on some open neighborhood W of the zero-section of $T(M)$ by the property that $\exp(X) = \gamma(1)$ where $X \in T_p(M)$ and γ is the geodesic so that $\gamma(0) = p$ and $\gamma'(0) = X$. The exponential map is equivariant in the sense that if $X, gX \in W$, then $\exp(gX) = g\ \exp(X)$. Let η be the orthogonal complement of $T(A)$ in $T(M)$, i.e., η is the normal bundle of A in M. Then one can find a smooth function

$$\varepsilon : A \to \mathbb{R}_{>0}$$

(constant if A is compact) so that

$$\exp : \mathring{D}_\varepsilon(\eta) \to M$$

is a smooth embedding onto a open neighborhood of A in M. The tubular neighborhood is then obtained using a fiber- and zero-section preserving diffeomorphism $E(\eta) \cong \mathring{D}_\varepsilon(\eta)$. □

Applying this to the submanifolds $\{x\}$ where $x \in M^G$, one gets

COROLLARY 1.1. *For a smooth action of a compact Lie group G on a manifold M, the fixed-point set M^G is a smooth submanifold.*

Let $x \in M$. The isotropy group G_x is closed in G, so is in fact a Lie subgroup. There is a canonical smooth structure on G/G_x so that $\pi : G \to G/G_x$ is a submersion. It is not difficult to show:

LEMMA 1.1. *Suppose a compact Lie group G acts smoothly on M. Let $x \in M$. Then the map $G/G_x \to M$, $g \mapsto gx$ is a smooth embedding. Hence the orbit Gx is a G-invariant submanifold of M.*

As a corollary of Theorem 1.2 and Lemma 1.1 one obtains:

THEOREM 1.3 (Slice theorem). *Suppose a compact Lie group G acts smoothly on a manifold M. Let $x \in M$. Then there is vector space V_x on which the isotropy group G_x acts linearly and a G-embedding*

$$G \times_{G_x} V_x \to M$$

onto an open set which sends $[g, 0]$ to gx.

For a right G-set A and a left G-set B, let $A \times_G B$ denote the quotient of $A \times B$ by the diagonal G-action. The image of $\{e\} \times V_x$ in M is called a *slice at x*. Here the representation $V_x = T_x(Gx)^\perp \subset T_x(M)$, where G acts via isometries of M. Then $G \times_{G_x} V$ is diffeomorphic to $T(Gx)^\perp$ and the map in the slice theorem is a G-invariant tubular neighborhood of the orbit Gx.

We now consider generalizations of the fact that M^G is a smooth submanifold. For a subgroup $H < G$, M^H need not be a manifold. However:

THEOREM 1.4 (Orbit theorem). *Suppose a compact Lie group acts smoothly on M.*

(i) *For any subgroup $H < G$,*

$$M_{(H)} = \{x \in M \mid H \text{ is conjugate to } G_x\}$$

is a smooth submanifold of G. The quotient map $\pi : M_{(H)} \to M_{(H)}/G$ is a smooth fiber bundle which can be identified with the bundle

$$G/H \times_{W(H)} (M_{(H)})^H \to (M_{(H)})^H/W(H),$$

where $W(H) = N(H)/H$ and $N(H)$ is the normalizer of H in G.

(ii) *Suppose M/G is connected, then there is an isotropy group H so that for all $x \in M$, H is conjugate to a subgroup of G_x. Moreover $M_{(H)}$ is open and dense in M and the quotient $M_{(H)}/G$ is connected.*

Since G-invariant submanifolds (e.g., Gx, M^G, $M_{(H)}$) have G-tubular neighborhoods, it behooves us to examine G-vector bundles. Recall that a finite-dimensional real representation E of a compact Lie group decomposes into a direct sum of irreducible representations. This decomposition is not canonical, but if one sums all isomorphic irreducible submodules of E, then one gets a canonical decomposition. The same thing works on the level of vector bundles.

Let $\mathrm{Irr}(G, \mathbb{R})$ be the set of isomorphism classes of finite-dimensional irreducible $\mathbb{R}G$-modules. For $[V] \in \mathrm{Irr}(G, \mathbb{R})$, let $D(V) = \mathrm{Hom}_{\mathbb{R}G}(V, V)$. Then $D(V)$ equals $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$.

PROPOSITION 1.2. *Let E be a G-vector bundle where G is a compact Lie group. Then*

$$\bigoplus_{[V]\in \mathrm{Irr}(G,\mathbb{R})} \mathrm{Hom}_{\mathbb{R}G}(V, E) \otimes_{D(V)} V \cong E,$$

where the map is $(f, v) \mapsto f(v)$. If $D(V) = \mathbb{C}$ (or $\mathbb{H}$) then the sub-bundle $\mathrm{Hom}_{\mathbb{R}G}(V, E) \otimes V$ admits a complex (or symplectic) structure.

COROLLARY 1.2. *Suppose $\mathbb{Z}/p$ acts smoothly on M with p prime.*

(i) *If p is odd, the normal bundle to the fixed set $M^{\mathbb{Z}/p} \subset M$ admits a complex structure.*

(ii) *If $p = 2$ and the action is orientation-preserving on an orientable manifold M, then $\dim M - \dim M^{\mathbb{Z}/2}$ is even.*

For homotopy theoretic information concerning a G-space, it is helpful to have the structure of a G-CW-complex.

DEFINITION 1.1. A *G-CW complex* is a G-space X together with a filtration

$$\phi = X_{-1} \subset X_0 \subset X_1 \subset X_2 \subset \cdots \subset X_n \subset \cdots \subset X = \bigcup_{n\geqslant 0} X_n$$

such that $X = \mathrm{colim}_{n\to\infty} X_n$ and for any $n \geqslant 0$ there is a pushout diagram

$$\begin{array}{ccc} \coprod_{i\in I_n} G/H_i \times \mathbb{S}^{n-1} & \longrightarrow & X_{n-1} \\ \downarrow & & \downarrow \\ \coprod_{i\in I_n} G/H_i \times D^n & \longrightarrow & X_n \end{array}$$

where $\{H_i\}_{i\in I_n}$ is a collection of subgroups of G.

Another point of view follows. A discrete group G acts cellularly on an ordinary CW complex X if for every $g \in G$ and for every open cell c of X, gc is an open cell of X and $gc = c$ implies that $g|_c = \mathrm{Id}$. Any cellular action on a CW complex X gives a G-CW-complex and conversely. From this point of view it is clear that if X is a G-CW-complex, so are X/G and X^H for all subgroups $H < G$.

Much of the elementary homotopy theory of CW-complexes remains valid for G-CW-complexes when G is discrete. For example, there is equivariant obstruction theory [26]. Note that specifying a G-map $G/H \times \mathbb{S}^{n-1} \to X_{n-1}$ is equivalent to specifying a map $\mathbb{S}^{n-1} \to X^H_{n-1}$. Using this observation, it is easy to show:

PROPOSITION 1.3 (Whitehead theorem). *Let $f : X \to Y$ be a G-map between G-CW-complexes. Then f is a G-homotopy equivalence (i.e., there is a G-map $g : Y \to X$ so that $f \circ g$ and $g \circ f$ are G-homotopic to the identity) if and only if $f^H : X^H \to Y^H$ induces an isomorphism on homotopy groups, for all subgroups H of G.*

A smooth G-manifold for G a finite group admits an equivariant triangulation, and hence the structure of a G-CW-complex [90]. The corresponding result for a smooth, proper action of a Lie group on a manifold appears in [93].

For a smooth G-manifold for a finite group G, much of the theory of differential topology goes through. For example, there are equivariant Morse functions and equivariant handle decompositions [168]. This leads to equivariant versions of the s-cobordism theorem, see [108, Section I.4.C] and the references therein. On the other hand, transversality fails equivariantly: consider the constant $\mathbb{Z}/2$-map $M \to \mathbb{R}$ from a manifold with a trivial $\mathbb{Z}/2$-action to the reals with the action $x \mapsto -x$; there is no homotopy to a map which is simultaneously equivariant and transverse.

1.5. *Change of category*

The subject of actions of groups on PL or topological manifolds differs from that of smooth actions on smooth manifolds. An action of a finite group on a topological manifold satisfies none of the regularity theorems of the previous section, and hence has been little studied. For example, one can suspend the involution on $\mathbb{S}^5$ with fixed set $L(d; 1, 1)$ to get an involution on $\mathbb{S}^6$ so that the fixed set (the suspension of the lens space) is not a manifold. Bing [22] constructed an involution on $\mathbb{S}^3$ with fixed set an Alexander horned sphere.

More typically studied are *topologically locally linear actions* of a compact Lie group on a topological manifold or *PL locally linear actions* of a finite group on a PL manifold.

By definition, these are manifolds with actions which satisfy the conclusion of the Slice Theorem 1.3. Such actions were called locally smooth in the older literature. For such actions the Orbit Theorem 1.4 remains valid; in particular the fixed set M^G is a submanifold. However, equivariant tubular neighborhoods and equivariant handlebodies need not exist. In fact, a locally linear action of a finite group on a closed manifold need not have the G-homotopy type of a finite G-CW-complex [142]. This makes the equivariant s-cobordism theorem [160,142] in this setting much more subtle; it requires methods from controlled topology. On the other hand, the theory of free actions of finite groups on closed manifolds parallels the smooth theory [102]. For general information on locally linear actions see [25,170].

1.6. *Remarks*

In this first section we have introduced basic objects, examples and questions associated to a topological transformation group. In the next section we will apply methods from algebraic topology to the study of group actions. As we shall see, these methods provide plenty of interesting invariants and techniques. After describing the main results obtained from this algebraic perspective, in Section 3 we will return to geometric questions. Having dealt with basic cohomological and homotopy-theoretic issues allows one to focus on the essential geometric problems by using methods such as surgery theory. Important examples such as the spherical space form problem will illustrate the success of this approach.

2. Cohomological methods in transformation groups

2.1. *Introduction*

In this section we will outline the important role played by cohomological methods in finite transformation groups. These ideas connect the geometry of group actions to accessible algebraic invariants of finite groups, hence propitiating a fruitful exchange of techniques and concepts, and expanding the relevance of finite transformation groups in other areas of mathematics. After outlining the basic tools in the subject, we will describe the most important results and then provide a selection of topics where these ideas and closely related notions can be applied. Although many results here apply equally well to compact Lie groups, for concreteness we will assume throughout that we are dealing with finite groups, unless stated otherwise. The texts by Allday and Puppe [10], Bredon [25] and tom Dieck [63] are recommended as background references.

To begin we recall a classical result due to Lefschetz: let X be a finite polyhedron and $f: X \to X$ a continuous mapping. The Lefschetz number $L(f)$ is defined as $L(f) = \sum_{i=0}^{\dim X} (-1)^i \operatorname{Tr} H_i(f)$, where $H_i(f): H_i(X; \mathbb{Q}) \to H_i(X; \mathbb{Q})$ is the map induced in rational homology. Lefschetz' fundamental fixed-point theorem asserts that if $L(f) \neq 0$, then f has a fixed point, i.e., an $x \in X$ such that $f(x) = x$. In particular this implies that if $G = \mathbb{Z}/n$ acts on an acyclic finite polyhedron X, then $X^G \neq \emptyset$. This result depends on the geometry of X as well as on the simple group-theoretic nature of $\mathbb{Z}/n$. How does this basic result generalize to more complicated groups? In the special case when G is a finite p-group (p a prime), P. Smith (see [25]) developed algebraic methods for producing fundamental fixed-point theorems of the type mentioned above. Rather than describe Smith Theory in its original form, we will outline the modern version as introduced by A. Borel in [23,24].

2.2. *Universal G-spaces and the Borel construction*

Denote a contractible free G-space by EG; such an object can be constructed functorially using joins, as was first done by Milnor in [118]. This space is often called a *universal G-space* and has the property that its singular chains are a free resolution of the trivial module over $\mathbb{Z}G$. A cellular model of EG can easily be constructed and from now on we will assume this condition. Now the quotient $BG = EG/G$ is a $K(G, 1)$, hence its cohomology coincides with the group cohomology $H^*(G, \mathbb{Z}) = \mathrm{Ext}^*_{\mathbb{Z}G}(\mathbb{Z}, \mathbb{Z})$. The space BG is also called the *classifying space* of G due to the fact that homotopy classes of maps into BG from a compact space Y will classify principal G-bundles over Y (a result due to Steenrod [159]). If X is a G-space, recall the *Borel construction* on X, defined as

$$X \times_G EG = (X \times EG)/G$$

where G acts diagonally (and freely) on the product $X \times EG$. If X is a point, we simply recover BG. If G is any non-trivial finite group, then EG is infinite-dimensional; hence if X is a G-CW complex, $X \times_G EG$ will be an infinite-dimensional CW complex. However, if G acts freely on X, then the Borel construction is homotopy equivalent to the orbit space X/G. The cohomology $H^*(X \times_G EG, \mathbb{Z})$ is often called the *equivariant cohomology of the G-space X*. In homological terms, this cohomology can be identified with the G-hypercohomology of the cellular cochains on X (see [41] or [33] for more on this). Let us assume from now on that X is a finite dimensional G-CW complex; although in many instances this condition is unnecessary, it does simplify many arguments without being too restrictive. The key fact associated to the object above is that the projection

$$X \times_G EG \to BG$$

is a fibration with fiber X, and hence we have a spectral sequence with

$$E_2^{p,q} = H^p\big(BG, \mathcal{H}^q(X; \mathbb{A})\big) \Rightarrow H^{p+q}(X \times_G EG; \mathbb{A}),$$

where $\mathbb{A}$ are the (possibly twisted) coefficients. Note that in addition G may act non-trivially on the cohomology of X. We are now in a position to explain the key results from Smith theory. Let $G = \mathbb{Z}/p$; then the inclusion of the fixed-point set $X^G \to X$ induces a map

$$i_G : X^G \times BG \to X \times_G EG$$

with the following property:

$$i_G^*: H^r(X \times_G EG; \mathbb{F}_p) \to H^r(X^G \times BG; \mathbb{F}_p)$$

is an isomorphism if $r > \dim X$, where $\mathbb{F}_p$ denotes the field with p elements. To prove this, we consider the G-pair (X, X^G) and the relative Borel construction $(X, X^G) \times_G EG = (X \times_G EG, X^G \times BG)$. The statement above is equivalent to showing that $H^r((X, X^G) \times_G EG; \mathbb{F}_p) = 0$ for r sufficiently large. However, this follows from the fact that the relative co-chain complex $C^*(X, X^G)$ is G-free, and hence the relative equivariant cohomology can be identified with the cohomology of the subcomplex of invariants, which vanishes above the dimension of X.

Now if X is mod p homologous to a point, then the spectral sequence collapses and looking at high dimensions we infer that X^G is mod p homologous to a point. If G is any finite p-group, it will always have a central subgroup of order p, hence using induction one can easily show

THEOREM 2.1 (Smith). *If a finite p-group G acts on a finite-dimensional complex X mod p homologous to a point, then X^G is non-empty and is also mod p homologous to a point.*

In contrast, it is possible to construct fixed-point free actions of $\mathbb{Z}/pq$ (where p, q are distinct primes) on $\mathbb{R}^n$ (see [25]). This indicates that p-groups play a distinguished part in the theory of group actions, analogous to the situation in group cohomology or representation theory.

If $G = \mathbb{Z}/p$ acts on $X = \mathbb{S}^n$ with a fixed point, the corresponding spectral sequence will also collapse. The key observation is that the existence of a fixed-point leads to a cross section for the bundle $X \times_G EG \to BG$, and hence no non-zero differentials can hit the cohomology of the base; as there are only two lines the spectral sequence must collapse. Using induction this yields

THEOREM 2.2 (Smith). *If a finite p-group G acts on a finite-dimensional complex X mod p homologous to a sphere with a fixed point, then X^G is also mod p homologous to a sphere.*

Much later, Lowell Jones (see [99]) proved a converse to Smith's theorem for actions on disks which goes as follows.

THEOREM 2.3 (Jones). *Any finite $\mathbb{F}_p$-acyclic complex is the fixed-point set of a $\mathbb{Z}/p$-action and thus of any finite p-group on some finite contractible complex.*

The spectral sequence used above can also be applied to prove the following basic result (see [24]).

THEOREM 2.4. *If G, a finite p-group, acts on a finite-dimensional complex Y, then*

$$\sum_{i=0}^{\dim Y} \dim H^i(Y;\mathbb{F}_p) \geqslant \sum_{i=0}^{\dim Y^G} \dim H^i\big(Y^G;\mathbb{F}_p\big).$$

Obvious examples such as actions on projective spaces can be analyzed using these techniques; for detailed applications we refer to [25, Chapter VII], which remains unsurpassed as a source of information on this topic. To give a flavour of the results there, we describe an important theorem due to Bredon. Let $P^h(n)$ denote a space such that its mod p cohomology is isomorphic to the ring $\mathbb{F}_p[a]/a^{h+1}$, where a is an element of dimension n.

THEOREM 2.5 (Bredon). *Suppose that p is prime and that $G = \mathbb{Z}/p$ acts on a finite-dimensional complex X with the mod p cohomology of $P^h(n)$. Then either $X^G = \emptyset$ or it is the disjoint union of components $F_1, \dots, F_k$ such that F_i is mod p cohomologous to $P^{h_i}(n_i)$, where $h + 1 = \sum_{i=1}^k (h_i + 1)$ and $n_i \leqslant n$. The number of components k is at most p. For p odd and $h \geqslant 2$, n and the n_i are all even. Moreover, if $n_i = n$ for some i, then the restriction $H^n(X;\mathbb{F}_p) \to H^n(F_i;\mathbb{F}_p)$ is an isomorphism.*

2.3. *Free group actions on spheres*

Next we consider applications to the spherical space form problem, namely what finite groups can act freely on a sphere? Let us assume that G does act freely on $\mathbb{S}^n$, then examining the spectral sequence as before we note that it must abut to the cohomology of an n-dimensional orbit space, hence the differential

$$d_{n+1} : H^k\big(G, H^n(\mathbb{S}^n;\mathbb{F}_p)\big) \to H^{k+n+1}(G;\mathbb{F}_p)$$

must be an isomorphism for k positive, and hence the mod p cohomology of G must be *periodic*. From the Kunneth formula, it follows that $G \neq (\mathbb{Z}/p)^n$ with $n > 1$; applying this to all the subgroups in G, we deduce that *every* abelian subgroup in G is cyclic, hence obtaining another classical result due to P. Smith.

THEOREM 2.6 (Smith). *If G acts freely on $\mathbb{S}^n$, then every abelian subgroup of G is cyclic.*

A finite group has all abelian subgroups cyclic if and only if its mod p cohomology is periodic for all p (see [41]). Groups which satisfy this condition have been classified and their cohomologies have been computed (see [6]). In this context, a natural question arises: does every periodic group act freely on a sphere? The answer is negative, as a consequence of a result due to Milnor [119]:

THEOREM 2.7 (Milnor). *If G acts freely on $\mathbb{S}^n$, then every element of order 2 in G must be central.*

Hence in particular the dihedral group D_{2p} cannot act freely on any sphere. Note that this result depends on the fact that the sphere is a manifold. However, such restrictions do not matter in the homotopy-theoretic context, as the following result due to Swan [161] shows:

THEOREM 2.8 (Swan). *Let G be a finite group with periodic cohomology; then it acts freely on a finite complex homotopy equivalent to a sphere.*

At this point the serious problem of realizing a geometric action must be addressed; this will be discussed at length in Section 3. As a preview we mention the theorem that a group G will act freely on some sphere if and only if every subgroup of order p^2 or $2p$ (p a prime) is cyclic; these are precisely the conditions found by Smith and Milnor.

Clearly the methods used for spheres can be adapted to look at general free actions, given some information on the cohomology of the group. The following example illustrates this: let G denote the semidirect product $\mathbb{Z}/p \times_T \mathbb{Z}/p-1$, where the generator of $\mathbb{Z}/p-1$ acts via the generator in the units of $\mathbb{Z}/p$. From this it is not hard to show that $H^*(G;\mathbb{Z}_{(p)}) \cong \mathbb{Z}_{(p)}[u]/pu$, where $u \in H^{2(p-1)}(G;\mathbb{Z}_{(p)})$, and $\mathbb{Z}_{(p)}$ denotes the integers localized at p. Now assume that G acts freely on a connected complex X, such that the action is *trivial on homology*. From the spectral sequence associated to this action we can infer the following: the dimension of X must be at least $2(p-1)-1$. If it were less, then no differential in the spectral sequence could hit the generator from the base; and hence $H^{2(p-1)}(X/G;\mathbb{Z}_{(p)}) \neq 0$, a contradiction.

2.4. *Actions of elementary abelian groups and the localization theorem*

Let us now assume that $G = (\mathbb{Z}/p)^r$, an elementary abelian p-group. Cohomological methods are extremely effective for studying actions of these groups. Perhaps the most important result is the celebrated "Localization Theorem" due to Borel and Quillen [140]. To state it we first recall that if $x \in H^1(G;\mathbb{F}_p)$ is non-zero, then its Bockstein $\beta(x)$ is a two-dimensional polynomial class. Let $0 \neq e \in H^{2(p^r-1)}(G;\mathbb{F}_p)$ denote the product of all the $\beta(y)$, as y ranges over non-zero elements in $H^1(G;\mathbb{F}_p)$.

THEOREM 2.9 (Borel and Quillen). *Let $G = (\mathbb{Z}/p)^r$ act on a finite-dimensional complex X. Then, if S is the multiplicative system of powers of e, the localized map induced by inclusions*

$$S^{-1}H^*(X \times_G EG;\mathbb{F}_p) \to S^{-1}H^*(X^G \times BG;\mathbb{F}_p)$$

is an isomorphism.

This result has substantial applications to the theory of finite transformation groups. Detailed results about fixed-point sets of actions on spheres, projective spaces, varieties, etc. follow from this, where in particular information about the ring structure of the fixed-point set can be provided. An excellent source of information on this is the text by Allday and Puppe [10]. An important element to note is that the action of the Steenrod algebra is an essential additional factor which can be used to understand the fixed-point set (see also [67]). Also one should keep in mind the obvious interplay between the E_2 term of the spectral sequence described previously and the information about the E_∞ term the localization theorem provides. Important results which should be mentioned are due to Hsiang [82] and Chang and Skelbred [42]. In particular we have the following fundamental result.

THEOREM 2.10 (Chang and Skelbred). *If $G = (\mathbb{Z}/p)^r$ and X is a finite-dimensional G-CW complex which is also a* mod p *Poincaré duality space, then each component F_i of X^G is also a* mod p *Poincaré duality space.*

For the case of actions of compact Lie groups, Atiyah and Bott [15] describe a De Rham version of the localization theorem, which is quite useful for studying questions in differential geometry and physics (see also [65]). There are also recent applications of localization techniques to problems in symplectic geometry, for example in [98].

2.5. *The structure of equivariant cohomology*

We now turn to describing qualitative aspects of equivariant cohomology which follow from isotropy and fixed point data. This was originally motivated by attempts to understand the asymptotic growth rate (Krull dimension) of the mod p cohomology of a finite group G. Atiyah and Swan conjectured that it should be precisely the the rank of G at p (i.e., the dimension of its largest p-elementary abelian subgroup). This result was in fact proved by Quillen [140] in his landmark work on cohomology of groups. First we need some notation. Denote by $\mathcal{A}_G$ the family of all elementary abelian p-subgroups in G, and by $\mathcal{A}_G(X)$ the ring of families $\{f_A : X^A \to H^*(A; \mathbb{F}_p)\}_{A \in \mathcal{A}_G}$ of locally constant functions compatible with respect to inclusion and conjugation. Consider the homomorphism $H^*(X \times_G EG; \mathbb{F}_p) \to \mathcal{A}_G(X)$ which associates to a class u the family $(\tilde{u}_A)$, where $(\tilde{u}_A)$ is the locally constant function whose value at x is the image of u under the map in equivariant cohomology associated to the inclusion $A \subset G$ and the map from a point to X with image $\{x\}$.

THEOREM 2.11 (Quillen). *If X is compact, then the homomorphism above is an F-isomorphism of rings, i.e., its kernel and cokernel are both nilpotent.*

The following two results follow from Quillen's work

PROPOSITION 2.1. *Let G act on a finite complex X and denote by $p(t)$ the Poincaré series for the mod p equivariant cohomology of X. Then $p(t)$ is a rational function of the form $z(t)/\prod_{i=1}^{n}(1-t^{2i})$, where $z(t) \in \mathbb{Z}[t]$, and the order of the pole of $p(t)$ at $t = 1$ is equal to the maximal rank of an isotropy subgroup of G.*

PROPOSITION 2.2. *If G is a finite group, then the map induced by restrictions*

$$H^*(G; \mathbb{F}_p) \to \lim_{A \in \mathcal{A}_G} H^*(A; \mathbb{F}_p)$$

is an F-isomorphism.

For example, if $G = \mathcal{S}_n$, the finite symmetric group, then the map above is actually an isomorphism for $p = 2$. We refer the reader to the original paper for complete details; it suffices to say that the proof requires a careful consideration of the Leray spectral sequence associated to the projection $X \times_G EG \to X/G$.

This result has many interesting consequences; here we shall mention that it was the starting point to the extensive current knowledge we have in the cohomology of finite groups (see [6]). An analogous theorem for modules has led to the theory of complexity and many connections with modular representations have been uncovered (see [37] and [19]).

EXAMPLE 2.1. The following simple example ties in many of the results we have discussed. Let $G = Q_8$, the quaternion group of order 8. Its mod 2 cohomology is given by (see [6,41])

$$H^*(G; \mathbb{F}_2) \cong \mathbb{F}_2[x_1, y_1, u_4]/x_1^2 + x_1 y_1 + y_1^2, x_1^2 y_1 + x_1 y_1^2.$$

Note that the asymptotic growth rate of this cohomology is precisely one, which corresponds to the fact that it is periodic. In addition every element of order 2 is central; in fact $Q_8 \subset \mathbb{S}^3$ and hence acts freely on it by translation. The class u_4 is polynomial, transgressing from the top-dimensional class in $\mathbb{S}^3$. In fact one can see that

$$H^*(G; \mathbb{F}_2)/(u_4) \cong H^*(\mathbb{S}^3/Q_8; \mathbb{F}_2)$$

which means that the classes $1, x_1, y_1, x_1 y_1, x_1^2, x_1 y_1^2$ represent a cohomology basis for the mod 2 cohomology of the 3-manifold $\mathbb{S}^3/Q_8$. The unique elementary abelian subgroup is the central $\mathbb{Z}/2$, and the four-dimensional class u_4 restricts to $e_1^4 \in H^4(\mathbb{Z}/2; \mathbb{F}_2)$, where e_1 is the one-dimensional polynomial generator. The other cohomology generators are nilpotent.

Another interesting group which acts freely on $\mathbb{S}^3$ is the binary icosahedral group B of order 120 (it is a double cover of the alternating group $\mathcal{A}_5$). In this case we have

$$H^*(B; \mathbb{F}_2) \cong \Lambda(x_3) \otimes \mathbb{F}_2[u_4],$$

where as before in the spectral sequence for the group action the top class in the sphere transgresses to u_4. From this we obtain $H^*(\mathbb{S}^3/B; \mathbb{F}_2) \cong \Lambda(x_3)$. This orbit space is the Poincaré sphere.

These examples illustrate how geometric information is encoded in the cohomology of a finite group, a notion which has interesting algebraic extensions (see [21]).

2.6. *Tate cohomology, exponents and group actions*

The cohomology of a finite group can always be computed using a free resolution of the trivial $\mathbb{Z}G$ module $\mathbb{Z}$. It is possible to splice such a resolution with its dual to obtain a *complete resolution* (see [6]), say $\mathcal{F}_*$, indexed over $\mathbb{Z}$, with the following properties:

(1) each $\mathcal{F}_i$ is free,

(2) $\mathcal{F}_*$ is acyclic and

(3) $\mathcal{F}_*$, $* \geqslant 0$, is a free resolution of $\mathbb{Z}$ in the usual sense.

Now let X be a finite-dimensional G-CW complex; in [162] Swan introduced the notion of equivariant Tate cohomology, defined as

$$\widehat{H}^i_G(X) = H^i\bigl(\mathrm{Hom}_G(\mathcal{F}_*; C^*(X))\bigr).$$

An important aspect of the theory is the existence of two spectral sequences abutting to the Tate cohomology, with respective E_1 and E_2 terms

$$E_1^{p,q} = \widehat{H}^q\bigl(G; C^p(X)\bigr)$$

and

$$E_2^{p,q} = \widehat{H}^p\bigl(G; H^q(X; \mathbb{Z})\bigr)$$

which arise from filtering the associated double complex in the two obvious ways. Using the fact that free modules are Tate-acyclic, the first spectral sequence can be used to show that $\widehat{H}^*_G(X) \equiv 0$ if and only if the G-action is free. More generally one can show that equivariant Tate cohomology depends only on the singular set of the action. In addition it is not hard to see that $\widehat{H}^*_G(X) \cong H^*(X \times_G EG; \mathbb{Z})$ for $* > \dim X$. Recent work has concentrated on giving a homotopy-theoretic definition of this concept and defining analogues in other theories (see [5,75]). This involves using a geometric construction of the transfer. Another important ingredient is the 'homotopy fixed point set' defined as $\mathrm{Map}_G(EG, X)$; in fact an analysis of the natural map $X^G = \mathrm{Map}_G(*, X) \to X^{hG}$ is central to many important results in equivariant stable homotopy.

Let A be a finite abelian group; we define its exponent $\exp(A)$ as the smallest integer $n > 0$ such that $n.a = 0$ for all $a \in A$. Using the transfer, it is elementary to verify that $|G|$ annihilates $\widehat{H}^*_G(X)$; hence exponents play a natural role in this theory. Assume that X is a connected, free G-CW complex. Now consider the $E_r^{0,0}$ terms in the second spectral sequence described above; the possible differentials involving it are of the form

$$E_{r+1}^{-r-1,r} \to E_{r+1}^{0,0} \to E_{r+2}^{0,0}$$

with $r = 1, 2, \ldots, N$, $N = \dim X$. From these sequences we obtain that $\exp E_{r+1}^{0,0}$ divides the product of $\exp E_{r+1}^{-r-1,r}$ and $\exp E_{r+2}^{0,0}$ and hence as $E_\infty^{0,0} = 0$, and $E_2^{0,0} = \widehat{H}^0(G; \mathbb{Z}) = \mathbb{Z}/|G|$, we obtain the following condition, first proved by Browder (see [28,1]): $|G|$ divides the product $\prod_{r=1}^{\dim X} \exp \widehat{H}^{-r-1}(G; H^r(X; \mathbb{Z}))$. We note the following important consequence of this fact.

THEOREM 2.12 (Browder). *If X is a connected, free $(\mathbb{Z}/p)^r$-CW complex, and if the action is trivial in homology, then the total number of dimensions $i > 0$ such that $H^i(X; \mathbb{Z}_{(p)}) \neq 0$ must be at least r.*

COROLLARY 2.1 (Carlsson). *If $(\mathbb{Z}/p)^r$ acts freely and cellularly on $(\mathbb{S}^n)^k$ with trivial action in homology, then $r \leqslant k$.*

This corollary, was proved by Carlsson [38] using different methods. In [4] the hypothesis of homological triviality was removed for odd primes and hence we have the generalization of Smith's result, namely

THEOREM 2.13 (Adem–Browder). *If p is an odd prime and $(\mathbb{Z}/p)^r$ acts freely on $(\mathbb{S}^n)^k$, then $r \leqslant k$.*

For $p = 2$ the same result will hold provided $n \neq 3, 7$. This is a Hopf invariant one restriction. The case $n = 1$ is due to Yalcin [175].

Another consequence of Browder's result concerns the exponents carried by the Chern classes of a faithful unitary representation of G.

COROLLARY 2.2. *Let $\rho : G \to U(n)$ denote a faithful unitary representation of a finite group G. Then $|G|$ must divide the product $\prod_{i=1}^{n} \exp(c_i(\rho))$.*

Using methods from representation theory, one can in fact show [2] that for $G = (\mathbb{Z}/p)^r$,

$$\exp \widehat{H}_G^*(X) = \exp \widehat{H}_G^0(X) = \max\{|G_x|,\ x \in X\}$$

hence in particular we obtain for any G

THEOREM 2.14. *The Krull dimension of $H^*(X \times_G EG; \mathbb{F}_p)$ is equal to the maximum value of $\log_p\{\exp \widehat{H}_E^0(X)\}$ as E ranges over all elementary abelian p subgroups of G.*

This shows the usefulness of equivariant Tate cohomology, as it will determine asymptotic cohomological information for ordinary equivariant cohomology from a single exponent.

In [29], Browder defined the degree of an action as follows. Let G act on a closed oriented manifold M^n preserving orientation, and let $j : M \to M \times_G EG$ denote the fiber inclusion. Then

$$\deg_G(M) = \left|H^n(M; \mathbb{Z})/\operatorname{im} j^*\right|.$$

This was independently defined by Gottlieb in [74]; they both show that if $G = (\mathbb{Z}/p)^r$, then $\log_p \deg_G(M)$ is equal to the co-rank of the largest isotropy subgroup in G. Note in particular that the action will have a fixed-point if and only if $\deg_G(M) = 1$. Using duality it is possible to prove their result from the previous theorem, we refer the reader to [2] for details.

2.7. *Acyclic complexes and the Conner conjecture*

If X is a G space and $H \subset G$ is a subgroup then a basic construction is the *transfer map* $C_*(X/G) \to C_*(X/H)$. By averaging on cochains it is elementary to construct such a map (see [33]) with the property that composed with the projection $X/H \to X/G$ the resulting map is multiplication by $[G : H]$ on $H^*(X/G, \mathbb{A})$, where $\mathbb{A}$ is any coefficient group. Note in particular that if $P = \mathrm{Syl}_p(G)$, we have an embedding $H^*(X/G; \mathbb{F}_p) \to H^*(X/P; \mathbb{F}_p)$. A basic result is

THEOREM 2.15. *If X is a finite-dimensional acyclic G-complex, then X/G is acyclic.*

From the above, to show that X/G is acyclic it suffices to show (for any prime p) that if X is mod p acyclic and $\mathbb{Z}/p$ acts on X, then $X/\mathbb{Z}/p$ is mod p acyclic. Consider the the mod p equivariant cohomology of the relative cochain complex for the pair $(X, X^{\mathbb{Z}/p})$; as it is free, we can identify it with the mod p cohomology of the quotient pair, $(X/\mathbb{Z}/p, X^{\mathbb{Z}/p})$. Now the E_2 term of the spectral sequence converging to this is of the form $H^p(G, H^q(X, X^{\mathbb{Z}/p}; \mathbb{F}_p))$; using the fact that the fixed-point set must be mod p acyclic (by Smith's theorem) we conclude that it must be identically zero and hence $H^*(X/\mathbb{Z}/p, \mathbb{F}_p) \cong H^*(X^{\mathbb{Z}/p}, \mathbb{F}_p)$ and so $X/\mathbb{Z}/p$ is mod p acyclic. Less obvious is the fact that if X is contractible, then so is X/G (see [63, p. 222]). The most general results along these lines are due to Oliver [129] who in particular settled a fundamental conjecture due to Conner for compact Lie groups.

THEOREM 2.16 (Oliver). *Any action of a compact Lie group on a Euclidean space has contractible orbit space.*

The main elements in the proof are geometric transfers and a careful analysis of the map $X \times_G EG \to X/G$ which we discussed previously. Oliver also proved some results about fixed-point sets of smooth actions on discs [128], extending a basic example due to Floyd–Richardson (see [25] for details) in a remarkable way.

We introduce a few group-theoretic concepts. Let $\mathcal{G}_p^q$ be the class of finite groups G with normal subgroups $P \lhd H \lhd G$ such that P is of p-power order, G/H is of q-power order and H/P is cyclic. Let

$$\mathcal{G}_p = \bigcup_q \mathcal{G}_p^q, \qquad \mathcal{G} = \bigcup_p \mathcal{G}_p.$$

We can now state

THEOREM 2.17 (Oliver). *A finite group G has a smooth fixed-point free action on a disk if and only if $G \notin \mathcal{G}$. In particular, any non-solvable group has a smooth fixed-point free action on a disk, and an abelian group has such an action if and only if it has three or more non-cyclic Sylow subgroups.*

COROLLARY 2.3. *The smallest abelian group with a smooth fixed-point free action on a disk is* $\mathbb{Z}/30 \oplus \mathbb{Z}/30$, *of order* 900. *The smallest group with such an action is the alternating group* $\mathcal{A}_5$ *of order* 60.

Note that $\mathcal{A}_5$ is precisely the group occurring in the Floyd–Richardson example. Oliver proved a more general version.

THEOREM 2.18. *For any finite group* G *not of prime power order, there is an integer* n_G *(the Oliver number) so that a finite CW-complex* K *is the fixed-point set of a* G*-action on some finite contractible complex if and only if* $\chi(K) \equiv 1 \pmod{n_G}$. *Furthermore, if* $\chi(K) \equiv 1 \pmod{n_G}$ *there is a smooth* G*-action on a disk with fixed-point set homotopy equivalent to* K.

Recently, Oliver [135] has returned to this problem and by analyzing G-vector bundles, has determined the possible fixed-point sets of smooth G-actions on some disk when G is not a p-group.

We now include a small selection of topics in finite transformation groups to illustrate the scope and diversity of the subject, as well as the significance of its applications. This is by no means a complete listing, but hopefully it will provide the reader with interesting examples and ideas.

2.8. *Subgroup complexes and homotopy approximations to classifying spaces*

Let G denote a finite group and consider $S_p(G)$, the partially ordered set of all non-trivial p-subgroups in G. G acts on this object via conjugation and hence on its geometric realization $|S_p(G)|$, which is obtained by associating an n-simplex to a chain of $n+1$ subgroups under inclusion. Hence we obtain a finite G-CW complex inherently associated to any finite group G. Similarly if $A_p(G)$ denotes the poset of non-trivial p-elementary abelian subgroups, $|A_p(G)|$ will also be a finite G-CW complex. These complexes were introduced by K. Brown and then studied by Quillen [141] in his foundational paper. He showed that these complexes have properties analogous to those of Tits Buildings for finite groups of Lie type. Moreover, these geometric objects associated to finite groups are of substantial interest to group theorists, as they seem to encode interesting properties of the group.

We now summarize basic properties of these G-spaces.

(1) $|S_p(G)|$ is G equivariantly homotopic to $|A_p(G)|$.

(2) For all p-subgroups $P \subset G$, the fixed point set $|S_p(G)|^P$ is contractible.

(3) There is an isomorphism

$$\widehat{H}^*(G; \mathbb{F}_p) \cong \widehat{H}^*_G(|S_p(G)|; \mathbb{F}_p)$$

(due to Brown [33]).

(4) In the mod p Leray spectral sequence for the map $|A_p(G)| \times_G EG \to |A_p(G)|/G$ we have that $E_2^{p,q} = 0$ for $p > 0$ and $E_2^{0,q} \cong H^q(G; \mathbb{F}_p)$. This means that $H^*(G; \mathbb{F}_p)$ can be computed from the cohomology of the normalizers of elementary abelian subgroups and their intersections (this is due to P. Webb, see [169] and [6]).

The following example illustrates the usefulness of these poset spaces.

EXAMPLE 2.2. Let $G = M_{11}$, the first Mathieu group. We have that $|A_2(G)|$ is a finite graph, with an action of G on it such that the quotient space is a single edge, with vertex stabilizers Σ_4 and $GL_2(\mathbb{F}_3)$ and edge stabilizer D_8 (dihedral group of order 8). From this information the cohomology of G can be computed (at $p = 2$), and we have (see [6])

$$H^*(G; \mathbb{F}_2) \cong \mathbb{F}_2[v_3, u_4](w_5)/w_5^2 + v_3^2 u_4.$$

Moreover, from the theory of trees we have a surjection

$$\Sigma_4 *_{D_8} GL_2(\mathbb{F}_3) \to G$$

which is in fact a mod 2 cohomology equivalence. Hence the poset space provides an interesting action which in turns leads to a 2-local model for the classifying space of a complicated (sporadic) simple group. More generally this technique can be used to show that if K is a finite group containing $(\mathbb{Z}/p)^2$ but not $(\mathbb{Z}/p)^3$, then at p the classifying space BK can be modelled by using a virtually free group arising from the geometry of the subgroup complex, which is a graph. We refer to [6] for more complicated instances of this phenomenon.

In a parallel development, important recent work in homotopy theory has focused on constructing 'homotopy models' for classifying spaces of compact Lie groups (see [94,95]). In particular the classifying spaces of centralizers of elementary abelian subgroups can be used to obtain such a model (again p-locally). This is related to cohomological results but has a deeper homotopy-theoretic content which we will not discuss here. We suggest the recent paper by Dwyer [66] for a thorough exposition of the homotopy decompositions of classifying spaces. Equivariant methods play an important part in the proofs.

We should also mention that if G is a perfect group, then the homotopy groups $\pi_n(BG^+)$ contain substantial geometric information, often related to group actions. Here BG^+ denotes Quillen's *plus construction* which is obtained from BG by attaching two and three dimensional cells and has the property of being simply connected, yet having the same homology as BG. We refer the interested reader to [6, Chapter IX], for details.

2.9. *Group actions and discrete groups*

An important application of finite transformation groups is to the cohomology of discrete groups of finite virtual cohomological dimension, as first suggested by Quillen in [140]. These are groups Γ which contain a *finite index* subgroup Γ' of finite cohomological dimension (i.e., with a finite-dimensional classifying space). Examples will include groups such as amalgamated products of finite groups, arithmetic groups, mapping class groups, etc. If for example $\Gamma \subset GL_n(\mathbb{R})$ is a discrete subgroup, then Γ will act on the symmetric space $GL_n(\mathbb{R})/K$ (K a maximal compact subgroup) with *finite* isotropy. Analogous models and their compactifications are the basic building blocks for approaching the cohomology of discrete groups.

More abstractly, using a simple coinduction construction due to Serre (see [33]), one can always build a finite dimensional Γ-CW complex X such that

(1) $X^H \neq \emptyset$ if and only if $H \subset \Gamma$ is finite,

(2) X^H is contractible for all H finite.

Now we can choose Γ' to be a normal subgroup of finite cohomological dimension and finite index in Γ. Hence the finite group $G = \Gamma/\Gamma'$ will act on the finite-dimensional space X/Γ', with isotropy subgroups corresponding to the finite subgroups in Γ. Moreover, it is not hard to see that for a finite subgroup $H \subset \Gamma$,

$$(X/\Gamma')^H \simeq \coprod_{(J)} B\big(N_\Gamma(J) \cap \Gamma'\big),$$

where J runs over all Γ'-conjugacy classes of finite subgroups of Γ mapping onto H via the projection $\Gamma \to G$ and $N_\Gamma(J)$ is the normalizer of J in Γ (see [33]).

We are therefore in an ideal situation to apply Smith theory to obtain a lower bound on the size of the cohomology of these discrete groups. To make it quite general, we assume given Γ of finite cohomological dimension and P a finite p-group of automorphisms for Γ. Let $\overline{\Gamma} = \Gamma \times_T P$, the semi-direct product; now Γ is a normal subgroup of finite index in this group. If we choose $J \subset \overline{\Gamma}$ a finite subgroup mapping onto P, let $C_\Gamma(J)$ denote its centralizer in Γ. Let $H^1(P, \Gamma)$ denote the usual non-abelian cohomology and finally denote by $\dim_{\mathbb{F}_p} H^*(Y)$ the total dimension of the homology $\sum H^i(Y; \mathbb{F}_p)$ for a finite-dimensional complex Y. We can now state (see [3]):

THEOREM 2.19. *If Γ is a discrete group of finite cohomological dimension, then for every finite p-group of automorphisms P of Γ we have*

$$\dim_{\mathbb{F}_p} H^*(\Gamma) \geqslant \sum_{J \in H^1(P,\Gamma)} \dim_{\mathbb{F}_p} H^*\big(C_\Gamma(J)\big)$$

and in particular

$$\dim_{\mathbb{F}_p} H^*(\Gamma) \geqslant \dim_{\mathbb{F}_p} H^*\big(\Gamma^P\big),$$

where $\Gamma^P \subset \Gamma$ is is the fixed subgroup under the automorphism group P.

As an application of this, we have that if $\Gamma_n(q) \subset SL_n(\mathbb{Z})$ denotes a level q (q prime) congruence subgroup, and if p is another prime, then

$$\dim_{\mathbb{F}_p} H^*\big(\Gamma(q)\big) \geqslant 2^{k(p-3)/2} \cdot \dim_{\mathbb{F}_p} H^*\big(\Gamma_t(q)\big),$$

where $n = k(p-1) + t$, $0 \leqslant t < p - 1$.

The summands in the general formula will represent ‘topological special cycles’ which in more geometric situation intersect to produce cohomology (see [150]). A result such as the above should be a basic tool for constructing non-trivial cohomology for discrete groups with symmetries; in fact groups such as the congruence subgroups will have many finite automorphisms and hence plentiful cohomology. Equivariant techniques should continue to be quite useful in producing non-trivial cohomology.

We should also mention that Brown [33] used equivariant methods to prove very striking results about Euler characteristics of discrete groups. The following is one of them. The group theoretic Euler characteristic of Γ (situation as in the beginning of this section) can be defined as $\chi(\Gamma) = \chi(\Gamma')/|G|$; one checks that it is indeed well-defined. Now let $n(\Gamma)$ denote the least common multiple of the orders of all finite subgroups in Γ. Serre conjectured and K. Brown proved that in fact $n(\Gamma) \cdot \chi(\Gamma) \in \mathbb{Z}$. This beautiful result furnishes information about the size of the finite subgroups in Γ, provided the Euler characteristic can be computed. In many instances this is the case; for example, $\chi(Sp_4(\mathbb{Z})) = -1/1440$, from which we deduce that $Sp_4(\mathbb{Z})$ has subgroups of order 32, 9 and 5. From a more elementary point of view, this result is simply a consequence of the basic fact that the least common multiple of the orders of the isotropy subgroups of a finite-dimensional G-complex Y (with homology of finite type) must yield an integer when multiplied by $\chi(Y)/|G|$.

2.10. *Equivariant K-theory*

After the usual cohomology of CW complexes was axiomatized by Eilenberg and Steenrod, the introduction of 'extraordinary' theories led to many important results in topology; specifically K-theory was an invaluable tool in solving a number of problems. Atiyah [13] introduced an equivariant version of K-theory whose main properties were developed by Segal [155] and Atiyah and Segal in [16]. We will provide the essential definitions and the main properties which make this a very useful device for studying finite group actions.

Equivariant complex K-theory is a cohomology theory constructed by considering equivariant vector bundles on G-spaces. Let X denote a finite G-CW complex, a G-vector bundle on X is a G-space E together with a G-map $p: E \to X$ such that

(i) $p: E \to X$ is a complex vector bundle on X,

(ii) for any $g \in G$ and $x \in X$, the group action $g: E_x \to E_{gx}$ is a homomorphism of vector spaces.

Assuming that G is a compact Lie group and X is a compact G-CW complex then the isomorphism classes of such bundles give rise to an associated Grothendieck group $K_G^0(X)$, which as in the non-equivariant case can be extended to a $\mathbb{Z}/2$ graded theory $K_G^*(X)$, the equivariant complex K-theory of X. An analogous theory exists for real vector bundles. We now summarize the basic properties of this theory:

(1) If X and Y are G-homotopy equivalent, then $K_G^*(X) \cong K_G^*(Y)$. However, in contrast to ordinary equivariant cohomology, an equivariant map $X \to Y$ inducing a homology equivalence does not necessarily induce an equivalence in equivariant K-theory (see [16]).

(2) $K_G^*(\{x_0\}) \cong R(G)$, the complex representation ring of G.

(3) Let $\mathcal{P} \subset R(G)$ denote a prime ideal with support a subgroup $S \subset G$ (in fact S is characterized as minimal among subgroups of G such that $\mathcal{P}$ is the inverse image of a prime of $R(S)$; if $\mathcal{P}$ is the ideal of characters vanishing at $g \in G$, then $S = \langle g \rangle$),

denote by $X^{(S)}$ the set of elements $x \in X$ such that S is conjugate to a subgroup of G_x; then we have the following localization theorem due to Segal:

$$K_G^*(X)_{\mathcal{P}} \cong K_G^*\big(X^{(S)}\big)_{\mathcal{P}}.$$

(4) If G is a finite group, then (see [62])

$$K_G^*(X) \otimes \mathbb{Q} \cong \bigoplus_{(g)} K^*\big(X^{(g)}/C_G(g)\big) \otimes \mathbb{Q},$$

where g varies over all conjugacy classes of elements in G. Using this it is possible to identify the Euler characteristic of $K_G^*(X) \otimes \mathbb{Q}$ with the so-called 'orbifold Euler characteristic' [78].

(5) (Completion theorem, [16])

$$K^*(X \times_G EG) \cong K_G^*(X)\widehat{\ },$$

where completion on the right is with respect to the augmentation ideal $I \subset R(G)$ and the module structure arises from the map induced by projection to a point. This is an important result, even for the case when X is a point; it implies that the K-theory of a classifying space can be computed from the completion of the complex representation ring.

We should mention that there is a spectral sequence for equivariant K-theory similar to the Leray spectral sequence discussed before for the projection from the Borel construction onto the orbit space, but which will involve the representation rings of the isotropy subgroups. These basic properties make equivariant K-theory a very useful tool for studying group actions, we refer to [16,25,63] for specific applications. The localization theorem ensures that it is particularly effective for actions of cyclic groups. Of course K-theory is also important in index theory [17].

2.11. *Equivariant stable homotopy theory*

Just as in the case of cohomology and K-theory, there is an equivariant version of homotopy theory. In its simplest setting, if G is a finite group and X, Y are finite G-CW complexes, then we consider G-homotopy classes of equivariant maps $f : X \to Y$, denoted $[X, Y]^G$. Such objects and the natural analogues of classical homotopy theoretic results have been studied by Bredon [26] and others, and there is a fairly comprehensive theory. In many instances results are reduced to ordinary homotopy theoretic questions on fixed-point sets, etc. Rather than dwell on this fairly well-understood topic, we will instead describe the basic notions and results in *equivariant stable homotopy theory*, which have had substantial impact in algebraic topology.

Let V denote a finite-dimensional real G-module and $\mathbb{S}^V$ its 1-point compactification. If X is a finite G-CW complex and Y an arbitrary one (both with fixed base points), we can define

$$\{X, Y\}^G = \lim_{U \in \mathcal{U}_G} \big[\mathbb{S}^U \wedge X, \mathbb{S}^U \wedge Y\big]^G,$$

where $\mathcal{U}_G$ is a countable direct sum of finite-dimensional $\mathbb{R}G$-modules so that every irreducible appears infinitely often and the limit is taken over the ordered set of all finite-dimensional G-subspaces of $\mathcal{U}_G$ under inclusion; and the maps in the directed system are induced by smashing with $\mathbb{S}^{U_1^\perp \cap U_2}$ and identifying $\mathbb{S}^{U_2}$ with $\mathbb{S}^{U_1^\perp \cap U_2} \wedge \mathbb{S}^{U_1}$, where $U_1 \subseteq U_2$. One checks that this is independent of $\mathcal{U}_G$ and identifications using the fact that the limit is attained, by an equivariant suspension theorem.

We can define $\pi_n^G(X) = \{\mathbb{S}^n, X\}^G$ and $\pi_G^n(X) = \{X, \mathbb{S}^n\}^G$, where X is required to be finite in the definition of π_G^n. The following summarizes the basic properties of these objects.

(1) (tom Dieck [63])

$$\pi_*^G(X) \cong \bigoplus_{(H)} \pi_*^s\left(EWH_+ \wedge_{WH} X^H\right),$$

where $WH = NH/H$ and the sum runs over all conjugacy classes of subgroups in G.

(2) $\pi_*^G(X)$ and $\pi_G^*(Y)$ are finitely generated for each value of $*$, if X is an arbitrary G-complex and Y is a finite G-complex.

(3) $\pi_0^G(\mathbb{S}^0) \cong A(G)$ as rings, where $A(G)$ is the Burnside ring of G (see [39]). Note that $\pi_G^{-*}(\mathbb{S}^0) \cong \pi_*^G(\mathbb{S}^0)$ is a module over $A(G)$.

Given the known facts about group cohomology and the complex K-theory of a finite group, it became apparent that the stable cohomotopy of BG_+ would be an object of central interest in algebraic topology. Segal conjectured that in dimension zero it should be isomorphic to the I-adic completion of the Burnside ring, an analogue of the completion theorem in K-theory (I the augmentation ideal in $A(G)$). This was eventually proved by G. Carlsson in his landmark 1984 paper (see [39]).

THEOREM 2.20 (Carlsson). *For G a finite group, the natural map $\widehat{\pi_G^*}(\mathbb{S}^0) \to \pi_s^*(BG_+)$ is an isomorphism, where $\widehat{\pi_G^*}(\mathbb{S}^0)$ denotes the completion of $\pi_G^*(\mathbb{S}^0)$ at the augmentation ideal in $A(G)$.*

A key ingredient in the proof is an application of Quillen's work on posets of subgroups to construct a G-homotopy equivalent model of the singular set of a G-complex X which admits a manageable filtration. The consequences of this theorem have permeated stable homotopy theory over the last decade and in particular provide an effective method for understanding the stable homotopy type (at p) for classifying spaces of finite groups (see [115]). For more information we recommend the survey by Carlsson [40] on equivariant stable homotopy theory.

This concludes the selected topics we have chosen to include to illustrate the relevance of methods from algebraic topology to finite transformation groups. Next we provide a short list of problems which are relevant to the material discussed in this section.

2.12. *Miscellaneous problems*

(1) Let G denote a finite group of rank n. Show that G acts freely on a finite-dimensional CW-complex homotopy equivalent to a product of n spheres $\mathbb{S}^{m_1} \times \cdots \times \mathbb{S}^{m_n}$.

(2) Prove that if $G = (\mathbb{Z}/p)^r$ acts freely on $X = \mathbb{S}^{m_1} \times \cdots \times \mathbb{S}^{m_n}$ then $r \leqslant n$.
(3) Show that if $(\mathbb{Z}/p)^r$ acts freely on a connected CW complex X, then

$$\sum_{i=0}^{\dim X} \dim_{\mathbb{F}_p} H^i(X; \mathbb{F}_p) \geqslant 2^r.$$

(4) Find a fixed integer N such that if G is any finite group with $H_i(G; \mathbb{Z}) = 0$ for $i = 1, \ldots, N$, then $G = \{1\}$.
(5) Calculate $K_G^*(|S_p(G)|)$ in representation-theoretic terms.
(6) Show that $|A_p(G)|$ is contractible if and only if G has a non-trivial normal p-subgroup.

REMARK 2.1. We have listed only a few, very specific problems which seem directly relevant to a number of questions in transformation groups. Problem (1) would be a generalization of Swan's result, and seems rather difficult. In [20], a solution was provided in the realm of projective kG chain complexes. Problem (2) has been around for a long time and again seems hard to approach. Problem (3) is a conjecture due to G. Carlsson; implies (2) and has an analogue for free chain complexes of finite type. Problem (4) has a direct bearing (via the methods in 2.9) on the problem of (given G) determining the minimal dimension of a finite, connected CW complex with a free and homologically trivial action of G. Problem (5) is a general formulation of a conjecture due to Alperin in representation theory, as described by Thevenaz [163]. Finally, Problem (6) is a conjecture due to Quillen [141] which has been of some interest in finite group theory (see [11]).

In this section we have attempted to summarize some of the basic techniques and results on the algebraic side of the theory of finite transformation groups. Our emphasis has been to make available the necessary definitions and ideas; additional details can be found in the references. It should however be clear that cohomological methods are a fundamentally useful device for studying transformation groups. In the next section we will consider the more geometric problem of actually constructing group actions when all algebraic restrictions are satisfied; as we will see, the combined approach can be quite effective but unfortunately also rather complicated.

3. Geometric methods in transformation groups

The subject of group actions on manifolds is diverse, and the techniques needed for future research seem quite unpredictable, hence we reverse our order of exposition in this section, and start with a discussion of five open problems, the solutions of which would lead to clear advances.

3.1. *Five conjectures*

(i) *Borel conjecture*: If a discrete group Γ acts freely and properly on contractible manifolds M and N with compact quotients, then the quotients are homeomorphic.

(ii) *Group actions on* $\mathbb{S}^3$ *are linear*: Any smooth action of a finite group on $\mathbb{S}^3$ is equivalent to a linear action.
(iii) *Hilbert–Smith Conjecture*: Any locally compact topological group acting effectively on a connected manifold is a Lie group.
(iv) *Actions on products of spheres*: If $(\mathbb{Z}/p)^r$ acts freely on $\mathbb{S}^{m_1} \times \cdots \times \mathbb{S}^{m_n}$, then $r \leqslant n$. More generally, what finite groups G act freely on a product of n spheres?
(v) *Asymmetrical manifolds*: There is a closed, simply-connected manifold which does not admit an effective action of a finite group.

3.1.1. *The Borel conjecture.* It may be a stretch to call the Borel conjecture a conjecture in transformation groups, but once one has done this, it has to be listed first, as it is one of the main principles of geometric topology. As such, it exerts its influence on transformation groups.

A space is *aspherical* if its universal cover is contractible. The Borel conjecture as stated is equivalent to the conjecture that any two closed, aspherical manifolds with isomorphic fundamental groups are homeomorphic. An aspherical manifold might arise in nature as a complete Riemannian manifold with non-positive sectional curvature or as $\Gamma \setminus G/K$ where Γ is a discrete, co-compact, subgroup of a Lie group G with a finite number of components and K is a maximal compact subgroup of G, however, the Borel conjecture is a general conjecture about topological manifolds. This is a very strong conjecture; in dimension 3 it implies the Poincaré conjecture, since if Σ^3 is a homotopy 3-sphere, the conclusion of the Borel conjecture applied to $T^3 \sharp \Sigma^3$ and $T^3 \sharp \mathbb{S}^3$ implies that $\Sigma^3 \cong \mathbb{S}^3$ by Milnor's prime decomposition of 3-manifolds [121]. Nonetheless, the conjecture has been proven in many cases: where one manifold is the n-torus T^n, $n \geqslant 4$ [71,89,102,166], or if one of the manifolds has dimension $\geqslant 5$ and admits a Riemannian metric of sectional curvature $K \leqslant 0$ [70]. In the study of the Borel conjecture in dimension 3, it is traditional to assume that both manifolds are irreducible, which means that any embedded 2-sphere bounds an embedded 3-ball. This assumption is made to avoid connected sum with a homotopy 3-sphere, and we will call the conjecture that homotopy equivalent, closed, irreducible, aspherical 3-manifolds are homeomorphic the *irreducible Borel conjecture*. The irreducible Borel conjecture has been proven when one of the manifolds is a torus [127], sufficiently large [164], Seifert fibered [154], and work continues in the hyperbolic case [72]. The irreducible Borel conjecture for general hyperbolic 3-manifolds and the Borel conjecture for hyperbolic 4-manifolds remains open.

What is the motivation for the Borel conjecture? First, from homotopy theory – any two aspherical complexes with isomorphic fundamental groups are homotopy equivalent. But the real motivation for Borel's conjecture (made by A. Borel in a coffee room conversation in 1953) was rigidity theory for discrete, co-compact subgroups of Lie groups, in particular the then recent results of Malcev [112] on nilpotent groups and Mostow [125] on solvable groups. Mostow showed that if Γ_1 and Γ_2 are discrete, co-compact subgroups of simply-connected solvable Lie groups G_1 and G_2 (necessarily homeomorphic to Euclidean space), and if $\Gamma_1 \cong \Gamma_2$, then the aspherical manifolds G_1/Γ_1 and G_2/Γ_2 are diffeomorphic. In the nilpotent case Malcev showed the stronger statement that there is an isomorphism $G_1 \to G_2$ which restricts to the given isomorphism $\Gamma_1 \to \Gamma_2$. Borel then speculated that while group theoretic rigidity sometimes failed, topological rigidity might always hold. Of course, such phenomena were known prior to the work of Malcev and Mostow. Bieberbach showed rigidity for crystallographic groups. On the other hand, failure of group theoretic rigidity was apparent from the existence of compact Riemann surfaces with the same genus and different conformal structures, i.e., there are discrete, co-compact subgroups of $SL_2(\mathbb{R})$ which are abstractly isomorphic, but there is no automorphism of $SL_2(\mathbb{R})$ which carries one to the other. The theory of group theoretic rigidity was investigated further by Mostow [126] and Margulis [113]. The subject of topological rigidity of group theoretic actions (as in Mostow's work on solvable groups) was pursued further by Raymond [147] and his collaborators.

We now discuss variants of the Borel conjecture. The Borel conjecture is not true in the smooth category: smoothing theory shows that T^n and $T^n \sharp \Sigma^n$, $n > 6$, are not diffeomorphic when Σ^n is an exotic sphere. The Borel conjecture is not true for open manifolds; there are contractible manifolds not homeomorphic to Euclidean space. This is shown by using the "fundamental group at infinity". In fact, Davis [57] constructed closed, aspherical manifolds which are not covered by Euclidean space. There are sharper forms of the Borel conjecture: *a homotopy equivalence between closed, aspherical manifolds is homotopic to a homeomorphism.* There is a reasonable version of the Borel conjecture for manifolds with boundary: *a homotopy equivalence between between compact, aspherical manifolds which is a homeomorphism on the boundary is homotopic, relative to the boundary, to a homeomorphism.*

What should be said for non-free actions? One might call the equivariant Borel conjecture the conjecture that if a discrete group Γ acts co-compactly on contractible manifolds X and Y so that the fixed point sets are empty for infinite subgroups of Γ and are contractible for finite subgroups of Γ, then X and Y are Γ-homeomorphic. This is motivated by the fact that they have the same Γ-homotopy type. Unfortunately, the equivariant Borel conjecture is not true, however, one can follow the philosophy of Weinberger [172] and take the success and failure of the equivariant Borel conjecture in particular cases as a guiding light for deeper investigation.

3.1.2. *Group actions on $\mathbb{S}^3$ are linear.* This is an old question, whose study breaks up into the cases of free and non-free actions. It seems likely that any solution requires geometric input. As is often the case in transformation groups on manifolds, the non-free actions are better understood. In particular, a key case is resolved. P.A. Smith showed that for a prime p, if $\mathbb{Z}/p$ acts smoothly, preserving orientation on $\mathbb{S}^3$ with a non-empty fixed point set, then the fixed set is an embedded circle. He conjectured that the fixed set is always unknotted. In [124], it was proven that such an action is equivariantly diffeomorphic to a linear action, giving the Smith conjecture. The proof, building on the work of Thurston, was the joint work of many mathematicians: Bass, Gordon, Litherland, Meeks, Morgan, Shalen, and Yau. The linearization question for general non-free actions is yet unresolved, waiting for a solution for the free case, but linearization results for many non-free actions are given in [124], and it has been shown that any smooth action of a finite group on $\mathbb{R}^3$ is equivalent to a linear action [105].

The case of free actions is still open, although there has been recent progress. The conjecture may be generalized: a closed 3-manifold with finite fundamental group is diffeomorphic to a linear spherical space form $\mathbb{S}^3/G$. For the trivial group, this is the Poincaré conjecture!

We note that it is not difficult to list the free, linear actions on $\mathbb{S}^3$; the fixed-point free subgroups of $SO(4)$ are given by Hopf [80], and then it is easy to give all free representations [174]. Work of Cartan and Eilenberg [41], Milnor [119], Lee [107], Milgram [116], and Madsen [109] gave restrictions on the possible finite fundamental groups of closed 3-manifolds. Hamilton [77] showed that a closed 3-manifold with a metric of positive Ricci curvature is diffeomorphic to a linear spherical space form.

Actions of finite groups on $\mathbb{S}^n$, $n \geqslant 4$, are reasonably well understood and need not be equivalent to linear actions. For surveys of non-free actions on $\mathbb{S}^n$, see [124] and [153]. For a survey of free actions on $\mathbb{S}^n$, see [52].

3.1.3. *Hilbert–Smith conjecture.* We take our discussion of the problem from the surveys of Raymond [146] and Yang [176]. We note at the outset that virtually no progress has been made on this conjecture during the last thirty years, so it may be the time for a fresh look.

The conjecture states that a locally compact topological group G acting effectively on a connected manifold M must be a Lie group. This is known in the following cases:

(i) $G = M$ and the group action is by multiplication. This is the famous result of Montgomery and Zippen which states that a manifold which admits a continuous group structure must be a Lie group.

(ii) M is a differentiable manifold and for all $g \in G$, multiplication by g gives a smooth map $M \to M$. In this case not only is G a Lie group, but the action is also smooth.

(iii) G is compact and every element of G is of finite order. The only such Lie groups are the finite groups.

An inverse limit of finite groups is totally disconnected, hence if the inverse limit is infinite, this gives an example of a compact group which is not a Lie group. The two most obvious examples of such are the infinite p-torus $\prod \mathbb{Z}/p$ for a prime p and the additive group $\widehat{\mathbb{Z}}_p$ of the p-adic integers. The infinite p-torus cannot act effectively on a manifold by result (iii) above. It is still an open question as to whether the p-adic integers can act. In fact, Yamabe has shown that every locally compact group has an open subgroup which is an inverse limit of Lie groups. Using this and (iii) above, one can shown if there is a counterexample to the Hilbert–Smith conjecture, then for some prime p, the p-adic integers act effectively on a manifold. Such an action would be strange indeed. If $\widehat{\mathbb{Z}}_p$ acts effectively on an n-dimensional manifold M, then $H^{n+2}(M/\widehat{\mathbb{Z}}_p; \mathbb{Z}) \neq 0$.

3.1.4. *Actions on a product of spheres.* This problem has been solved when the number of spheres is one; there is the result of Madsen, Thomas and Wall [111] which states that a finite group G acts freely on some sphere if and only if G has no non-cyclic abelian subgroups and no dihedral subgroups. When the number of spheres is greater than one, we discussed algebraic work in Section 2, but little geometric work has been done on this problem (but see [79,132], and [53]). It is evident that *any* finite group will act freely on a product of spheres. Simply take an element $g \in G$, make it act by rotation on an odd sphere and then induce up this action to an action of G on a product of spheres on which $\langle g \rangle$ still acts freely. Taking products over all elements $g \in G$ provides a product of spheres with a free G-action. The main problem which remains unsettled is to show that the number of spheres with any given free G-action will bound the rank of the elementary abelian subgroups in the group (see 2.12). On the *constructive* side, the following questions remain unanswered except in some special cases: if G is a finite group of rank $k > 1$ (rank is defined in terms of the maximal r_p, taken over all subgroups $(\mathbb{Z}/p)^{r_p}$), does G act freely on a finite dimensional CW-complex homotopy equivalent to a product of k spheres? If so, does G act freely on a product of k spheres? In recent work Adem and Smith (see [7]) showed that a finite p-group P acts freely on a finite complex $X \simeq \mathbb{S}^n \times \mathbb{S}^m$ if and only if P does not contain a subgroup isomorphic to $(\mathbb{Z}/p)^3$ and constructed actions of rank two simple groups (such as A_5 and $SL_3(\mathbb{F}_2)$) on homotopy products of two spheres.

We should mention that certain group-theoretic conditions can be used to produce the required free group actions. For example, if G is a finite 2-group of rank k satisfying Milnor's condition (i.e., every element of order 2 is central) then it will act freely on $(\mathbb{S}^{(|G|/2)-1})^k$. The action is built by inducing up sign representations on k elements of order 2 which span the unique central elementary abelian subgroup in G and then taking their product. More generally it is possible to use this approach to construct actions of arbitrary 2-groups on products of spheres with maximal isotropy of rank equal to the co-rank in G of the largest central elementary abelian subgroup. How to build a free action on a larger product from this object is still unknown. In the context of representation theory the work of U. Ray (see [145]) is also relevant here. She proves that if G is a finite group acting freely on a product of spheres arising from G-representations, then the only possible non-abelian composition factors of G are the alternating groups A_5 and A_6.

3.1.5. *Asymmetrical manifolds.* This problem is not as central as the other problems, but it does point out how little we know about group actions on manifolds not having "obvious" symmetries or manifolds closely resembling such. Presumably the asymmetrical manifold is the generic case (but don't ask what is precisely meant by that!) In the non-simply-connected case, asymmetrical manifolds were first constructed in [48].

3.2. *Examples and techniques*

3.2.1. *Non-linear similarity.* A fascinating chapter in the study of transformation groups is topological versus linear similarity. Two linear transformations $T, T' : V \to V$ of a finite-dimensional real vector space are *topologically similar* if $T' = hTh^{-1}$ for some homeomorphism $h : V \to V$. Elementary arguments (see [104]) show that if T and T' are topologically similar, then there are decompositions

$$V = V_f \oplus V_\infty, \qquad V = V'_f \oplus V'_\infty$$

invariant under T and T', respectively, such that $T|_{V_f}$ and $T'|_{V'_f}$ have finite (and equal) orders and are topologically similar, while $T|_{V_\infty}$ and $T'|_{V'_\infty}$ are linearly similar. Thus one may as well assume that T and T' have finite order. It was conjectured that topologically similar implies linearly similar, but this was disproved by Cappell and Shaneson [34] in 1981 using techniques from surgery theory. For $V = \mathbb{R}^9$ and for every $q > 1$, they constructed topologically similar T and T' of order $4q$ which are not linearly similar.

This problem is connected to many others in transformation groups. One first generalizes the problem; two finitely generated $\mathbb{R}G$-modules V and V' for a finite group G are *topologically similar* if there is an equivariant homeomorphism $h: V \to V'$.[1] (Note that if h is differentiable at the origin, then the differential $dh_0: V \to V'$ gives a linear similarity.) The modules V and V' are isomorphic to ones where G acts orthogonally; we assume the actions are orthogonal hereafter. Hence the actions restrict to the unit spheres. The G-spaces $S(V)$ and $S(V')$ are *topologically similar* if they are equivariantly homeomorphic. If $S(V)$ and $S(V')$ are topologically similar then so are V and V' (radially extend the homeomorphism) and conversely, if V and V' are topologically similar then $S(V \oplus \mathbb{R})$ and $S(V' \oplus \mathbb{R})$ are topologically similar (one-point compactify). Thus Cappell and Shaneson also constructed examples of non-linearly similar actions on spheres. If the actions on the spheres are free, then Whitehead–Reidemeister–De Rham torsion considerations (see the references in [44]), show that topologically similar actions on $S(V)$ and $S(V')$ are linearly similar. De Rham showed for general linear actions that if the spheres $S(V)$ and $S(V')$ are equivariantly smooth or PL-homeomorphic, then the representations are linearly similar, once again by torsion considerations (see the references in [108]). The fact that non-linear similarities exist implies that equivariant simple homotopy type is not a homeomorphism invariant.

Much of the analysis of the non-linear similarity problem stems from the following observation [35].

LEMMA 3.1. *V and V' are topologically similar if and only if $S(V)$ and $S(V')$ are G-h-cobordant.*

This means that there is a locally linear G-manifold W with boundary $S(V) \coprod S(V')$ so that there are equivariant, orbit-type preserving, strong deformation retracts of W onto $S(V)$ and onto $S(V')$. As part of the definition of G-h-cobordant, we also require that there is an inverse G-h-cobordism $-W$ so that $(-W) \cup_{S(V')} W$ and $W \cup_{S(V)} (-W)$ are G-homeomorphic rel ∂ to $S(V) \times I$ and $S(V') \times I$, respectively.

SKETCH OF PROOF. If $h: V \to V'$ is a topological similarity, then by re-scaling one may assume $h(D(V)) \subset \operatorname{int} D(V')$. Let $W = D(V') - \operatorname{int} D(V)$.

Conversely if W is a G-h-cobordism then

$$\begin{aligned} &\cdots \cup \big((-W) \cup W\big) \cup \big((-W) \cup W\big) \cup \cdots \\ &\quad \cong \cdots \cup (-W) \cup \big(W \cup (-W)\big) \cup (-W) \cup \cdots. \end{aligned}$$

Thus $S(V) \times \mathbb{R} \cong S(V') \times \mathbb{R}$. By adding on a point $\{+\infty\}$ to each to compactify one of the ends, we obtain our desired topological similarity. □

[1] This notion is connected with foundational issues in study of locally linear actions. If G acts locally linearly on a topological manifold M, then every point $x \in M$ has a neighborhood of the form $G \times_{G_x} V$. The $\mathbb{R}G_x$-module V is only determined up to topological similarity.

Construction of non-linearity similarities then proceeds using surgery theory and delicate algebraic number theory. Later approaches use the theory of topological equivariant h-cobordisms [35] or bounded methods [76].

The clearest positive result is:

THEOREM 3.1. *If G has odd order, then topologically similar representations are linearly similar.*

This "Odd Order Theorem" has four different proofs. The result is due independently to to Hsiang and Pardon [88] using stratified pseudo-isotopy theory and lower K-groups and to Madsen and Rothenberg [110] using equivariant smoothing theory. Later proofs were given by Rothenberg and Weinberger [151] using Lipschitz analysis and by Hambleton and Pedersen [76] using bounded methods.

Finally we would like to mention a related problem discussed by Shaneson in [156]. An G-action on a sphere Σ is said to be of *Smith type* if for all subgroups H, the fixed set Σ^H of H is either discrete or connected.

Given a smooth G-action on Σ of Smith type so that $\Sigma^G = \{x, y\}$, are the representations of G on the tangent spaces $T_x\Sigma$ and $T_y\Sigma$ linearly similar?

This question and the similarity question have the same flavor. Given an action on Σ as above, then Σ minus invariant open disks surrounding x and y gives a candidate for a G-h-cobordism between G-spheres. Conversely, given a G-h-cobordism W between G-spheres, then a candidate action on a sphere Σ is the end-point compactification of

$$\cdots \cup (-W) \cup W \cup (-W) \cup W \cup \cdots.$$

The study of the non-linear similarity question and the Smith question run parallel; however with the Smith question there are additional considerations of smoothness. Cappell and Shaneson answered Smith's question in the negative, while Sanchez [152] showed that the answer is affirmative for actions of groups of odd order. Cappell and Shaneson modify Smith's question to the conjecture that the tangent space representations are topologically similar. Earlier, Petrie constructed two fixed-point G-actions on spheres, with a 2-point fixed set and non-linearly similar tangent space representations, however, which were not of Smith type; we discuss these actions in a later section.

3.2.2. *Propagation of group actions.* Smith theory gives a connection between a group action and homological information at the order of the group. For example, if a group G acts semifreely (i.e., freely away from the fixed set) on a disk or a sphere, then Smith theory shows that the fixed set is a mod $|G|$-disk or sphere. Given a manifold with a group action, propagation is a systematic method for producing group actions on manifolds homologically resembling the given one. Three prototypical questions are:

(i) Given a mod $|G|$-homology sphere Σ and a free G-action on a sphere, does there exist a free G-action on Σ?
(ii) What are the fixed sets of semifree actions on disks?
(iii) What are the fixed sets of semifree actions on spheres?

Many mathematicians have worked on related ideas; we refer to [36,55,170], and [56] for references to original sources. These ideas were pioneered by Jones [99] and were taken farthest by S. Weinberger and his collaborators.

For the rest of this section, let q denote the order of the finite group G.

DEFINITION 3.1. A G-action on Y *propagates* across a map $f: X \to Y$ if there exists a G-action on X and an equivariant map homotopic to f.

Similarly given a G-action on X one can talk about propagation across a map $f: X \to Y$.

PROPOSITION 3.1. *Let $f: X \to Y$ be a map between simply-connected CW-complexes with $H_*(f; \mathbb{Z}/q) = 0$. Suppose G acts freely on Y, trivially on $H_*(Y; \mathbb{Z}[1/q])$. Then there is a complex X' and a homotopy equivalence $h: X' \to X$ so that the G-action propagates across $f \circ h$. Furthermore the homotopy type of X'/G is uniquely determined.*

We will sketch a proof of the above proposition, to illustrate the homotopy theoretic importance of homological triviality. For a set of primes P and a CW-complex X, there is a localization map $X \to X_{(P)}$, unique up to homotopy, inducing an isomorphism on π_1 and a localization of the higher homotopy groups. For an integer n, we use the notation $X_{1/n}$ and $X_{(n)}$ to mean invert the primes dividing n and not dividing n, respectively. The following lemma is due to Weinberger and is accomplished via a plus construction.

LEMMA 3.2. *Let Z be a CW-complex with finite fundamental group G. Then G acts trivially on $H_*(\widetilde{Z}; \mathbb{Z}[1/q])$ if and only if*

$$Z_{1/q} \simeq \widetilde{Z}_{1/q} \times BG.$$

PROOF OF PROPOSITION 3.1. Let X'/G be the homotopy pullback of

$$\begin{array}{ccc} & & X_{1/q} \\ & & \downarrow \\ (Y/G)_{(q)} & \longrightarrow & X_{(0)} \times BG. \end{array}$$

The propagation question can often (in fact, usually) be solved when

$$f: X \to Y$$

is a map between manifolds with $H_*(f; \mathbb{Z}/q) = 0$ and when the G-action is trivial on $H_*(\ ; \mathbb{Z}[1/q])$. However the general answer [56], phrased in terms of K- and L-theory and associated algebraic number theory, is too technical to state here. We give a few examples. □

EXAMPLE 1. Let Σ be a closed, oriented manifold of dimension n, $n \geqslant 5$, n odd, having the $\mathbb{Z}/q$-homology of the sphere. Then any free G-action on $\mathbb{S}^n$ can be propagated across any map $f: \Sigma \to \mathbb{S}^n$ whose degree is congruent to 1 modulo q.

EXAMPLE 2. Let $\Omega^k \subset D^n$ be a smoothly, properly embedded mod p-homology disk (p prime) so that there is a $\mathbb{Z}/p$-action on the normal bundle with fixed set Ω^k. (This happens when the normal bundle admits a complex structure.) Let N be a closed tubular neighborhood. Then the action propagates across the inclusion $\partial N \to D^n - \operatorname{int} N$. Thus there is a smooth, semifree $\mathbb{Z}/p$-action on D^n with fixed set Ω^k.

We next give three theorems giving answers to our three prototypical questions, but note there are many other variations of answers in the literature.

THEOREM 3.2 (Davis–Weinberger [55]). *Let G be a finite group. Let Σ^n (n odd, $n \geqslant 5$) be a closed, simply-connected manifold with $H_*(\Sigma^n; \mathbb{Z}/|G|) \cong H_*(\mathbb{S}^n; \mathbb{Z}/|G|)$. Then G acts freely on $\mathbb{S}^n$ if and only if G acts freely on Σ^n, trivially on homology.*

This produces free actions on manifolds which have no apparent symmetries, as long as they homologically resemble the sphere.

THEOREM 3.3 (Jones, Assadi–Browder, Weinberger [170]). *Let $\Omega^k \subset D^n$ be a proper, smooth embedding with $n - k$ even and greater than 2. Then Ω^k is the fixed set of a semifree orientation-preserving G-action on the disk if and only if*

(i) $\widetilde{H}_*(\Omega^k; \mathbb{Z}/|G|) = 0$,

(ii) $\sum (-1)^i [\widetilde{H}_i(\Omega^k; \mathbb{Z})] = 0 \in \widetilde{K}_0(\mathbb{Z}G)$,

(iii) *the normal bundle of Ω^k admits a semifree orientation-preserving G-action with fixed set Ω^k.*

Here $\widetilde{K}_0(\mathbb{Z}G)$ is the Grothendieck group of finitely generated projective $\mathbb{Z}G$-modules modulo the subgroup generated by free modules. For a finitely generated $\mathbb{Z}G$-module M of finite homological dimension, $[M] \in \widetilde{K}_0(\mathbb{Z}G)$ denotes the Euler characteristic of a projective resolution. In the above theorem, the $\widetilde{K}_0$ obstruction vanishes for G cyclic.

THEOREM 3.4 (Weinberger [171]). *Let Σ^k be a PL-locally flat submanifold of $\mathbb{S}^n$ with $n - k$ even and greater than 2. Then Σ^k is the fixed set of a semifree orientation-preserving G-action on $\mathbb{S}^n$ if and only if Σ^k is a $\mathbb{Z}/|G|$-homology sphere, $\mathbb{S}^{n-k-1}$ admits a free linear G-action, and certain purely algebraically describable conditions hold for the torsion in the homology of Σ.*

One condition is condition (ii) from the previous theorem, but there are further conditions involving the Swan subgroup of L-theory.

Recently Chase [43] has made some progress on propagation in the non-homologically trivial case, but with the presence of addition geometric hypotheses. For example, he has shown the following.

THEOREM 3.5. *A simply-connected* mod 2 *homology sphere Σ^{2k}, $2k \geqslant 6$, has a free involution if and only if there exists an orientation-reversing homeomorphism $L: \Sigma \to \Sigma$ so that $L \circ L$ acts unipotently on homology.*

3.2.3. *Equivariant surgery.* Surgery theory is the primary tool for classification of manifolds and for the study of transformation groups. It is discussed in more detail in Section 3.3. One classically applies surgery theory to transformation groups by doing surgery to the free part of a group action. But in this section we briefly mention a more general type of surgery theory, equivariant surgery. Its development is beset with difficulties arising from a lack of equivariant transversality and embedding theory. The solution to the embedding difficulties is to assume the "gap hypothesis" – that $\dim M_{(H)} \geqslant 5$ for all $H \subset G$ and $\dim M_{(K)} \geqslant 2 \dim M_{(H)} + 1$ for all $K \subset H \subset G$, but this assumption is not very appetizing.

Classical surgery theory studies both the uniqueness problem of classifying manifolds up to homeomorphism and the existence problem of determining when a space has the homotopy type of a manifold, however, equivariant surgery theory bifurcates. The equivariant uniqueness question is studied through the Browder–Quinn isovariant theory [32], which is a special case of Weinberger's theory of surgery on stratified spaces [172]. The equivariant surgery to study the existence question was developed by Petrie [139] and was applied and extended by several others, including Dovermann and Dovermann–Schultz. Two of the successes of Petrie's theory were one-fixed point actions on spheres and Smith equivalence of representations and we will discuss these results.

A smooth G-action with one fixed point on $\mathbb{S}^n$ gives a fixed point free action on D^n (delete an open equivariant neighborhood of the fixed point), and such an action in turn gives a fixed point free action on $\mathbb{R}^n$ (delete the boundary of the disk). So we first discuss the easier problems of constructing fixed point free actions on Euclidean space and the closed disk. The first examples of fixed point free actions of finite groups on Euclidean space and the closed disk were due to Floyd and Floyd–Richardson, respectively, and examples are discussed in [25]. The characterization of groups which can so act was accomplished by Edmonds and Lee [68] and Oliver [128] in the two cases. The first example of a fixed point free action of a finite group on a sphere was due to E. Stein, and the characterization of which groups can act freely without fixed points on some sphere was an application of equivariant surgery due to Petrie [138].

For a finite group G, two real representations V and W are said to be *Smith equivalent* if there is a smooth G-action on a sphere Σ with fixed point set $\{x, y\}$ so that $V \cong T_x \Sigma$ and $W \cong T_y \Sigma$. The first result is along these lines is the following theorem, proven using elliptic differential operators, along with some algebraic number theory.

THEOREM 3.6 (Atiyah–Bott [14], Milnor [122]). *If a compact Lie group G acts smoothly and semifreely on a sphere with two fixed points, then the representations at the two fixed points are linearly isomorphic.*

Using equivariant surgery, Petrie and others constructed many examples of smooth actions of finite groups on spheres with two fixed points, but whose representations are not isomorphic (see [139] and the articles in [153]).

3.3. *Free actions on spheres*

The techniques we would like to introduce are the theory of the finiteness obstruction and K_0, simple homotopy theory and K_1, and surgery theory. Rather than introduce these topics abstractly, we would like to introduce these through a concrete geometric situation, the study of topological spherical space forms, manifolds whose universal cover is a sphere.

3.3.1. *Existence: homotopy theoretic techniques.* The existence question is given a finite group G, can it act freely on $\mathbb{S}^n$, and the uniqueness question is what is the classification of manifolds with fundamental group G and universal cover $\mathbb{S}^n$. We first discuss the existence question. We will take n to be odd, since by the Lefschetz fixed point theorem, the only group which can act freely on an even-dimensional sphere is the cyclic group of order 2.

If G acts freely on $\mathbb{S}^n$, then by considering the spectral sequence of the fibration $\mathbb{S}^n \to \mathbb{S}^n/G \to BG$, it is easy to show that $H^{n+1}(G, \mathbb{Z}) \cong \mathbb{Z}/|G|$ and that for any additive generator $\alpha \in H^{n+1}(G, \mathbb{Z})$, that $\cup\alpha : \widehat{H}^i(G; M) \to \widehat{H}^{i+n+1}(G; M)$ is an isomorphism for all i and for all $\mathbb{Z}G$-modules M. ($\widehat{H}$ is Tate cohomology.) Thus G cannot have a subgroup of the form $\mathbb{Z}/p \times \mathbb{Z}/p$. There is a converse.

THEOREM 3.7 (Artin–Tate [41, Chapter XII]). *Let G be a finite group. The following are equivalent:*

(i) *All abelian subgroups are cyclic.*
(ii) *Every Sylow p-subgroup is cyclic or generalized quaternionic.*
(iii) *For some n, $H^{n+1}(G, \mathbb{Z}) \cong \mathbb{Z}/|G|$.*
(iv) *For some n, there is an element $\alpha \in H^{n+1}(G, \mathbb{Z})$, so that $\cup\alpha : \widehat{H}^i(G; M) \xrightarrow{\sim} \widehat{H}^{i+n+1}(G; M)$ for all i and for all $\mathbb{Z}G$-modules M.*

A group satisfying the above conditions is said to be *periodic* and if $H^{n+1}(G, \mathbb{Z}) \cong \mathbb{Z}/|G|$ then G is said to have *period* $n + 1$. The periodic groups have been classified and fall into six families (see, e.g., [52]). With regard to (iii) and (iv), for any finite group G and for any $\alpha \in H^{n+1}(G, \mathbb{Z})$, one can show that $\cup\alpha$ is an isomorphism if and only if α is an additive generator of $H^{n+1}(G, \mathbb{Z})$ and $H^{n+1}(G, \mathbb{Z}) \cong \mathbb{Z}/|G|$.

For a group G of period $n + 1$, does G act freely on $\mathbb{S}^n$? Not in general, but it does up to homotopy, as we shall see shortly. More precisely, we will show there exists a *Swan complex of dimension n*, an n-dimensional CW-complex X with $\pi_1 X = G$ and $\widetilde{X} \simeq \mathbb{S}^n$. It is *polarized* if one fixes the identification of the fundamental group with G and fixes the orientation, i.e., the homotopy class of the homotopy equivalence $\widetilde{X} \to \mathbb{S}^n$. By a *periodic projective resolution of period* $n + 1$, we mean an exact sequence of $\mathbb{Z}G$-modules

$$0 \to \mathbb{Z} \to P_n \to \cdots \to P_1 \to P_0 \to \mathbb{Z} \to 0,$$

where the $\mathbb{Z}$'s have trivial G-actions, and the P_i's are projective over $\mathbb{Z}G$. An example of such is $0 \to r\mathbb{Z} \to C_*(\widetilde{X}) \to \mathbb{Z} \to 0$, where X is an n-dimensional Swan complex. By splicing periodic projective resolutions using the composite $P_0 \to \mathbb{Z} \to P_n$, one can form a projective $\mathbb{Z}G$-resolution of $\mathbb{Z}$, or a complete resolution in the sense of Tate cohomology. It follows that G has period $n + 1$. In fact, by mapping a projective resolution of $\mathbb{Z}$ to a periodic projective resolution, one defines the k-invariant $k \in H^{n+1}(G, \mathbb{Z})$ of the periodic projective resolution. Cup product with k induces periodicity, and hence k is a generator of the cyclic group $H^{n+1}(G, \mathbb{Z})$. The k-invariant determines the polarized homotopy type of a Swan complex and the chain homotopy type of a projective periodic resolution. The following theorem is due to Swan [161].

THEOREM 3.8. *The following are equivalent:*

(i) *G has period $n+1$.*

(ii) *There is a projective periodic resolution G of period $n+1$, where the P_i's are finitely generated.*

(iii) *There is a projective periodic resolution G of period $n+1$, where the P_i's are free.*

(iv) *There is an n-dimensional CW-complex X with $\pi_1 X = G$ and $\widetilde{X} \simeq \mathbb{S}^n$.*

Furthermore given any generator $\alpha \in H^{n+1}(G, \mathbb{Z})$, one can construct the resolutions in (ii) *and* (iii) *and the complex in* (iv) *with k-invariant α.*

DISCUSSION OF PROOF. The implication (i) $\Rightarrow$ (ii) is the most difficult, although purely homological. We refer the reader to [167]. Here Wall shows that given a generator $\alpha \in H^{n+1}(G, \mathbb{Z})$ and an exact sequence

$$P_n \to P_{n-1} \to \cdots \to P_0 \to \mathbb{Z} \to 0,$$

with the P_i's projective, one can find a periodic projective resolution

$$0 \to \mathbb{Z} \to P_n' \to P_{n-1}' \oplus P_{n-1} \to \cdots \to P_0 \to \mathbb{Z} \to 0$$

with α as k-invariant.

For (ii) $\Rightarrow$ (iii), given a periodic resolution with the P_i's finitely generated projective, add on $Q_0 \xrightarrow{\mathrm{Id}} Q_0$ (in degrees 1 and 0) where $F_0 = P_0 \oplus Q_0$ is free. Continuing inductively, one obtains

$$0 \to \mathbb{Z} \to P_n' \to F_{n-1} \to \cdots \to F_0 \to \mathbb{Z} \to 0$$

with the F_i's finitely generated free. Choose a complement Q_n' so that $P_n' \oplus Q_n' = F_n$ is free. Next we use the Eilenberg swindle

$$\begin{aligned} P_n' \oplus (F_n)^\infty &= P_n' \oplus (Q_n' \oplus P_n') \oplus (Q_n' \oplus P_n') \oplus \cdots \\ &\cong (P_n' \oplus Q_n') \oplus (P_n' \oplus Q_n') \oplus \cdots = (F_n)^\infty. \end{aligned}$$

Thus we can add on $(F_n)^\infty \xrightarrow{\mathrm{Id}} (F_n)^\infty$ (in degrees n and $n-1$) to the above periodic projective resolution to obtain a periodic free resolution.

(iii) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (iii) we have already discussed.

It remains to show ((i), (ii) and (iii)) $\Rightarrow$ (iv). Build a $K(G, 1)$ and let Y be its $(n-1)$-skeleton. Then for any generator $\alpha \in H^{n+1}(G, \mathbb{Z})$, we may find a periodic free resolution

$$0 \to \mathbb{Z} \to F_n \xrightarrow{\partial} F_{n-1} \oplus C_{n-1}(\widetilde{Y}) \to \cdots \to C_0(\widetilde{Y}) \to \mathbb{Z} \to 0$$

with k-invariant α. Then let $X = Y \cup (\bigvee \mathbb{S}^{n-1}) \cup (\bigcup e_n)$ where there is a sphere for each $\mathbb{Z}G$-basis element of F_{n-1} and an n-cell for each $\mathbb{Z}G$-basis element of F_n. The n-cells are attached by using the Hurewicz isomorphism $\pi_{n-1}(Y \vee (\bigvee \mathbb{S}^{n-1})) \cong \partial(F_n)$. $\square$

The natural question now is whether the P_i's in a periodic projective resolution can be taken to be simultaneously free and finitely generated, or, equivalently, whether the Swan complex can be taken to be finite. The answer to this question is very subtle, and historically was one of the motivations for algebraic K-theory.

Let use review the construction of the Swan complex X in the proof of Theorem 3.8. Starting with $Y = K(G,1)^{n-1}$, one can build a periodic projective resolution

$$0 \to \mathbb{Z} \to P_n \xrightarrow{\partial} F_{n-1} \oplus C_{n-1}(\widetilde{Y}) \to \cdots \to C_0(\widetilde{Y}) \to \mathbb{Z} \to 0,$$

where Y is a finite CW-complex of dimension $n-1$, F_{n-1} is finitely generated free, and P_n is finitely generated projective. If P_n were stably free (i.e., the direct sum of P_n and a f.g. free module is free), then we don't need the Eilenberg swindle; we could add on the free module in dimensions n and $n-1$ and construct (via the Hurewicz theorem) a finite Swan complex. There is a converse.

LEMMA 3.3. *Let* $\{Q_n \to \cdots \to Q_0\}$ *and* $\{Q'_n \to \cdots \to Q'_0\}$ *be chain homotopy equivalent chain complexes of projective modules over a ring R. Then*

$$Q_n \oplus Q'_{n-1} \oplus Q_{n-2} \oplus \cdots \cong Q'_n \oplus Q_{n-1} \oplus Q'_{n-2} \oplus \cdots.$$

COROLLARY 3.1. *The Swan complex X has the homotopy type of a finite complex if and only if* P_n *is stably free.*

PROOF OF LEMMA 3.3. First prove it when one complex is zero by induction on n. Next prove it in general by applying the acyclic case to the algebraic mapping cone of the chain homotopy equivalence. □

This result leads to the notion of the *reduced projective class group* $\widetilde{K}_0(R)$ *of a ring R*. Elements are represented by $[P]$ where P is a finitely generated projective R-module. Here $[P] = [Q]$ if and only if $P \oplus R^m \cong Q \oplus R^n$ for some m and n. This is a classical notion; if R is a Dedekind domain, then $\widetilde{K}_0(R)$ can be identified with the ideal class group of R (see [123]). The above corollary shows that a Swan complex X defines an element $[X] \in \widetilde{K}_0(\mathbb{Z}G)$ (represented by P_n) which vanishes if and only if X has the homotopy type of a finite complex.

Next we analyze what happens to the finiteness obstruction if the k-invariant is changed. Suppose X' and X are Swan complexes. Then there is a map $X' \to X$ of degree d, and the k-invariants satisfy $k' = dk$, with $(d, |G|) = 1$. Define $P_d = \ker(\varepsilon)$, where $\varepsilon : \mathbb{Z}G \to \mathbb{Z}/d$ is defined by $\varepsilon(\sum n_g g) = \sum n_g$. Then P_d is projective (in fact, $P_d \oplus P_e$ is free if $de \equiv 1 \pmod{|G|}$). The finiteness obstructions satisfy

$$[X'] = [X] + [P_d].$$

Swan [161] defined what is now called the *Swan subgroup* of $\widetilde{K}_0(\mathbb{Z}G)$ as $T(G) = \{[P_d] \mid (d, |G|) = 1\}$. For a group G having period $n+1$, one defines the *Swan finiteness obstruction* $\sigma_{n+1}(G) \in \widetilde{K}_0(\mathbb{Z}G)/T(G)$ by setting $\sigma_{n+1}(G) = [X]$ for any Swan complex X

of dimension n and fundamental group G. Then $\sigma_{n+1}(G) = 0$ if and only if there exists a *finite* Swan complex of dimension n and fundamental group G.

Here are three fascinating examples (see [52] for more details).

(i) (Swan [161]) There is a finite 3-dimensional Swan complex X for the dihedral group of order 6 (so $\sigma_4(D_6) = 0$). By Milnor [119], X does not have the homotopy type of a manifold.

(i) (Swan [161], Martinet [114]) There is a 3-dimensional Swan complex with fundamental group the quaternion group of order 8 which does not have the homotopy type of a *finite* complex, and hence does not have the homotopy type of a closed manifold. (There is a 3-dimensional quaternionic space form, thus $\sigma_4(Q_8) = 0$ but $T(Q_8) \neq 0$.)

(iii) (Davis [50], Milgram [116]) Let $G = \mathbb{Z}/3 \times_T Q_{16}$ be a semidirect product where the element of order 8 in the quaternion group of order 16 acts non-trivially on $\mathbb{Z}/3$. G has period 4 but there is no *finite* 3-dimensional Swan complex, and hence no closed 3-manifold with fundamental group G (cf. [107]) (and so $\sigma_4(G) \neq 0$). The above is the smallest group satisfying this and is due to Davis; Milgram was the first to find examples with non-zero finiteness obstruction.

Given the above one might wonder for which periodic groups there is actually a finite Swan complex, and what its dimension is. From the algebraic reduction outlined above it follows that if $H^*(G, \mathbb{Z})$ has period d, then there exists an integer M such that G acts freely on a $(kd - 1)$-dimensional homotopy sphere. However it turns out that it suffices to take $k = 2$ (see [167, Corollary 12.6]). The proof of this fact is computational and goes through the list of periodic groups. Hence the minimal group of period four mentioned above which does not act freely on a homotopy 3-sphere will indeed act freely on a finite homotopy 7-sphere. The work of Swan was generalized by Wall to the theory of the Wall finiteness obstruction, which we now describe.

DEFINITION 3.2. A CW-complex X is *finitely dominated* if there is a finite CW-complex X_f and maps $i : X \to X_f$ and $r : X_f \to X$ so that $r \circ i \simeq \mathrm{Id}_X$.

This is similar to saying a module is finitely generated projective if and only if it is the retract of a finitely generated free module.

LEMMA 3.4. *A Swan complex X is finitely dominated.*

PROOF. By the proof of Theorem 3.8 we may assume X is homotopy equivalent to $Z = Y \cup (\bigvee \mathbb{S}^{n-1}) \cup (\bigcup e^n)$ where Y is a finite $(n-1)$-dimensional CW-complex. Furthermore

$$C_*(\widetilde{Z}) = F_n^\infty \oplus P_n \to F_n^\infty \oplus C_{n-1}(\widetilde{Y}) \to \cdots \to C_0(\widetilde{Y}),$$

where $F_n = P_n \oplus Q_n$ is a finitely generated free module and

$$0 \to \mathbb{Z} \to P_n \xrightarrow{\partial} C_{n-1}(\widetilde{Y}) \to \cdots \to C_0(\widetilde{Y}) \to \mathbb{Z} \to 0$$

is a periodic projective resolution. Then construct $X_f = Y \cup (\bigcup e^n)$ with

$$C_*(X_f) = P_n \oplus Q_n \xrightarrow{\partial\oplus 0} C_{n-1}(\widetilde{Y}) \to \cdots \to C_0(\widetilde{Y}),$$

using the Hurewicz isomorphism $\pi_{n-1} Y \cong (\partial \oplus 0)(P_n \oplus Q_n)$ to attach the n-cells. There are obvious chain maps $i_* : C_*(\widetilde{X}) \to C_*(\widetilde{X}_f)$ and $r_* : C_*(\widetilde{X}_f) \to C_*(\widetilde{X})$ so that $r_* \circ i_*$ is chain homotopic to the identity. Use the relative Hurewicz theorem to extend the inclusions $Y \to X_f$ and $Y \to X$ to the desired maps $i : Z \to X_f$ and $r : X_f \to Z$. □

Wall [165] generalized this.

PROPOSITION 3.2. *A connected CW-complex X is finitely dominated if and only if $\pi_1 X$ is finitely presented and $C_*(\widetilde{X})$ is chain homotopy equivalent to a complex $P_k \to \cdots \to P_0$ of finitely generated projective $\mathbb{Z}[\pi_1 X]$-modules.*

DEFINITION 3.3. For a finitely dominated connected CW-complex X, the *Wall finiteness obstruction* $[X] \in \widetilde{K}_0(\mathbb{Z}\pi_1 X)$ is defined by $[X] = \Sigma(-1)^i [P_i]$ where $C_*(\widetilde{X})$ is chain homotopy equivalent to $\{P_k \to \cdots \to P_0\}$, a chain complex of finitely generated projective $\mathbb{Z}[\pi_1 X]$-modules.

By Lemma 3.3, $[X]$ is well-defined.

THEOREM 3.9 (Wall [165]). *Let X be a finitely dominated connected CW-complex. Then $[X] = 0$ if and only if X is homotopy equivalent to a finite CW-complex.*

3.3.2. *Uniqueness: homotopy theoretic techniques.* We now turn to the uniqueness question. The general question is the classification up to homeomorphism of manifolds covered by a sphere; we consider only a very special case, the classification of classical lens spaces $L = L(k; i_1, \dots, i_n)$. For references on this material see Cohen [44] or Milnor [122]. Recall that L has fundamental group $\mathbb{Z}/k$, its universal cover is $\mathbb{S}^{2n-1} \subset \mathbb{C}^n$, and the integers $i_1, \dots, i_n$ are relatively prime to k; they give the rotations in the complex factors of $\mathbb{C}^n$. It is easy to see that the map $[z_1, \dots, z_n] \mapsto [z_1^{i_1}, \dots, z_n^{i_n}]$ gives a map $L(k; 1, \dots, 1) \to L(k; i_1, \dots, i_n)$ of degree $\prod_{j=1}^n i_j$. The following proposition follows from the earlier discussion of k-invariants.

PROPOSITION 3.3. *Two lens spaces $L(k; i_1, \dots, i_n)$ and $L(k; i_1', \dots, i_n')$ are homotopy equivalent if and only if $i_1 \cdots i_n \equiv \pm a^n i_1' \cdots i_n' \pmod k$ for some integer a.*

The plus or minus corresponds with the choice of orientation and the factor of a^n corresponds with the identification of the fundamental group of the lens space with $\mathbb{Z}/k$.

There are some obvious diffeomorphisms between lens spaces. For example, if $(i_1, \dots, i_n)$ considered modulo k is a permutation of $(i_1', \dots, i_n')$, then permuting the complex coordinates gives a diffeomorphism from $L(k; i_1, \dots, i_n)$ to $L(k; i_1', \dots, i_n')$. Similarly, mapping $[z_1, \dots, z_j, \dots, z_n] \mapsto [z_1, \dots, \bar{z}_j, \dots, z_n]$ gives a diffeomorphism from

$L(k; i_1, \ldots, i_j, \ldots, i_n)$ to $L(k; i_1, \ldots, -i_j, \ldots, i_n)$. Finally, if a is relatively prime to k, the map $[z_1, \ldots, z_n] \mapsto [\zeta^a z_1, \ldots, \zeta^a z_n]$ gives a diffeomorphism from $L(k; i_1, \ldots, i_n)$ to $L(k; ai_1, \ldots, ai_n)$. Lens spaces exhibit rigidity; all diffeomorphisms are generated by the above three types.

THEOREM 3.10. *Let $L(k; i_1, \ldots, i_n)$ and $L' = L(k; i'_1, \ldots, i'_n)$ be two lens spaces. The following are equivalent.*

(i) *L and L' are isometric, where they are given the Riemannian metrics coming from being covered by the round sphere.*
(ii) *L and L' are diffeomorphic.*
(iii) *L and L' are homeomorphic.*
(iv) *L and L' have the same simple homotopy type.*
(v) *There are numbers $\varepsilon_1, \ldots, \varepsilon_n \in \{1, -1\}$ and a number a relatively prime to k, so that $(i_1, \ldots, i_n)$ is a permutation of $(a\varepsilon_1 i_1, \ldots, a\varepsilon_n i_n)$ modulo k.*

The implications (v) $\Rightarrow$ (i), (i) $\Rightarrow$ (ii), and (ii) $\Rightarrow$ (iii) are easy, and (iii) $\Rightarrow$ (iv) follows from a theorem of Chapman which states that homeomorphic finite CW-complexes have the same simple homotopy type. To proceed farther, one must introduce the torsion and the Whitehead group.For a ring R, embed $GL_n(R)$ in $GL_{n+1}(R)$ by $A \mapsto \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}$. Let $GL(R) = \bigcup_{n=1}^{\infty} GL_n(R)$. Define

$$K_1(R) = GL(R)/\big[GL(R), GL(R)\big]$$

and $\widetilde{K}_1(R) = K_1(R)/(-1)$ where $(-1) \in GL_1(R)$. If R is a commutative ring, the determinant gives a split surjection $\det : K_1(R) \to R^*$, and if R is a field, this is an isomorphism.

A *based R-module* is a free R-module with a specified basis. A chain complex C over a ring R is *based* if each C_i is based, *finite* if $\bigoplus_i C_i$ is a finitely generated R-module, and *acyclic* if the its homology is zero. We now assume, for the sake of simplicity of exposition, that R has the property that $R^n \cong R^m$ implies that $n = m$ (e.g., a group ring $\mathbb{Z}\pi$ has this property since it maps epimorphically to $\mathbb{Z}$). An isomorphism $f : M \to M'$ between based R-modules determines an element $[f] \in \widetilde{K}_1(R)$. Given an chain isomorphism $f : C \to C'$ between finite, based chain complexes, define the *torsion of f* by

$$\tau(f) = \prod [f_i : C_i \to C'_i]^{(-1)^i} \in \widetilde{K}_1(R).$$

We next indicate the definition of the torsion $\tau(C)$ of a finite, based, acyclic chain complex C. Implicit in the proof of Lemma 3.3 is an algorithm for computing it.

PROPOSITION 3.4. *Let $\mathcal{C}$ be the class of finite, acyclic, based chain complexes over R. Then there is a unique map $\mathcal{C} \to \widetilde{K}_1(R)$, $C \mapsto \tau(C)$ satisfying the following axioms:*

(i) *If $f : C \to C'$ is a chain isomorphism where $C, C' \in \mathcal{C}$, then $\tau(C') = \tau(f)\tau(C)$.*
(ii) *$\tau(C \oplus C') = \tau(C)\tau(C')$.*
(iii) *$\tau(0 \to C_n \xrightarrow{d} C_{n-1} \to 0) = (-1)^{n-1}[d]$.*

We use this to define the torsion of a homotopy equivalence. If $f : C \to C'$ is a chain homotopy equivalence between finite, based R-chain complexes, define $\tau(f) = \tau(C(f))$ where $C(f)$ is the algebraic mapping cone of f. The reader should check that when f is a chain isomorphism, our two definitions of $\tau(f)$ agree. With somewhat more difficulty, the reader can verify that axiom 1 above holds for a chain homotopy equivalence $f : C \to C'$ between acyclic, finite, based complexes.

Let X be a finite, connected CW-complex with fundamental group π. For every cell in X, choose a cell in the universal cover $\widetilde{X}$ lying above it, and an orientation for that cell. Then $C(\widetilde{X})$ is a finite, based $\mathbb{Z}\pi$-chain complex. Accounting for the ambiguity in the choice of basis, define the *Whitehead group* $Wh(\pi) = K_1(\mathbb{Z}\pi)/\{\pm g\}_{g \in \pi}$ where $\{\pm g\}$ refers to the image of a one-by-one matrix. For a homotopy equivalence $h : X \to Y$, where X and Y are finite, connected CW-complexes with fundamental group π, define the *torsion of* h, $\tau(h) \in Wh(\pi)$ by $\tau(h) = \tau(\widetilde{h} : C(\widetilde{X}) \to C(\widetilde{Y}))$. A homotopy equivalence h is a *simple homotopy equivalence* if $\tau(h) = 0$. There is a geometric interpretation: h is a simple homotopy equivalence if and only if X can be obtained from Y via a sequence of elementary expansions and collapses; see [44]. A compact manifold has a canonical simple homotopy type; for smooth manifolds this follows from triangulating the manifold [122] and for topological manifolds this is more difficult [102].

Given a homotopy equivalence between lens spaces

$$h : L = L(k; i_1, \ldots, i_n) \longrightarrow L' = L(k; i'_1, \ldots, i'_n)$$

it is possible to compute its torsion in terms of $(i_1, \ldots, i_n)$ and $(i'_1, \ldots, i'_n)$. We will do something technically easier, which still leads to a proof of of Theorem 3.10. Let $T \in \mathbb{Z}/k$ denote the generator used to define the lens spaces and let $\Sigma = 1 + T + \cdots + T^{k-1} \in \mathbb{Z}[\mathbb{Z}/k]$ denote the norm element. There is a decomposition of rings

$$\mathbb{Q}[\mathbb{Z}/k] \cong \mathbb{Q} \times \Lambda,$$

where $\Lambda = \mathbb{Q}[\mathbb{Z}/k]/\Sigma$.

DEFINITION 3.4. If L is a finite complex with fundamental group $\mathbb{Z}/k = \langle T \rangle$, so that T acts trivially on $H_*(\widetilde{L}; \mathbb{Q})$, define the *Reidemeister torsion*

$$\Delta(L) = \tau\big(C(L; \Lambda)\big) \in K_1(\Lambda)/\{\pm T\}.$$

Finally, we can outline the proof of (iv) implies (v) in Theorem 3.10. Let $h : L \to L'$ be a simple homotopy equivalence of lens spaces. Then

$$\tau\big(C(L; \Lambda)\big) = \tau\big(C(L'; \Lambda)\big) \in K_1(\Lambda)/\{\pm T\}.$$

We now wish to compute both sides. Implicit in the definition of a lens space is an identification of the fundamental group with $\mathbb{Z}/k$. By perhaps replacing $L' = L(k; i'_1, \ldots, i'_n)$ by the diffeomorphic space $L(k; ai'_1, \ldots, ai'_n)$, we assume that h induces the identity on the fundamental group.

For $\widetilde{L}(k; i_1)) = \mathbb{S}^1$ it is easy to see that

$$C(\widetilde{L}(k; i_1)) \cong (\mathbb{Z}[\mathbb{Z}/k] \xrightarrow{\cdot g^{j_1}-1} \mathbb{Z}[\mathbb{Z}/k]),$$

where $i_1 j_1 \equiv 1 \pmod{k}$. The decomposition

$$\widetilde{L} = \mathbb{S}^{2n-1} = \mathbb{S}^1 * \cdots * \mathbb{S}^1 = \widetilde{L}(k; i_1) * \cdots * \widetilde{L}(k; i_n)$$

gives $\widetilde{L}$ a CW-structure, and

$$\begin{aligned} C(\widetilde{L}) \cong \mathbb{Z}[\mathbb{Z}/k] &\xrightarrow{\cdot g^{j_n}-1} \mathbb{Z}[\mathbb{Z}/k] \xrightarrow{\cdot \Sigma} \cdots \\ &\xrightarrow{\cdot g^{j_2}-1} \mathbb{Z}[\mathbb{Z}/k] \xrightarrow{\cdot \Sigma} \mathbb{Z}[\mathbb{Z}/k] \xrightarrow{\cdot g^{j_1}-1} \mathbb{Z}[\mathbb{Z}/k]. \end{aligned}$$

After tensoring with Λ the complex is acyclic and $\cdot \Sigma$ is zero, so the other maps must be isomorphisms, and

$$\tau(C(L; \Lambda)) = \prod_{\ell=1}^{n} g^{j_\ell} - 1.$$

To complete the argument, some (not so sophisticated) algebraic number theory comes in. $\Lambda \cong \prod_{d|k} \mathbb{Q}[\zeta_d]$ where the product is over all divisor of n greater than 1, and ζ_d is a primitive dth root of 1. One then considers the quotient of the torsion of L and the torsion of L' as units in the corresponding number rings, and the Franz independence lemma (whose proof is similar to that of the Dirichlet unit theorem) says that the only way these units can be roots of unity is if the conditions of part (v) of Theorem 3.10 are satisfied. This gives the classification of classical lens spaces.

3.3.3. *General remarks.* The reason why we are spending so much time on spherical space forms is that this problem represents a paradigm for the construction and classification of finite group actions on manifolds. The existence discussion above moved from algebraic information to geometric information at the level of CW-complexes; this passage is accomplished via the Hurewicz theorem. The finiteness obstruction in K_0 measures when a finite CW-complex can be obtained. The torsion in K_1 allows the discussion of simple homotopy theory and allows the classification of classical lens spaces. It also plays a key role in transformation groups via the s-cobordism theorem.

What is missing? Two things – at least – the transition to non-free actions and the transition to manifold theory. The generalization of the homotopy theoretic techniques to the non-free equivariant case is in reasonable shape. For example, de Rham used torsion to classify linear but non-free, actions of finite groups on spheres. The general theory of the equivariant finiteness obstruction, equivariant simple homotopy theory, and the equivariant s-cobordism theorem is worked out (see [108]) at least for the PL and smooth cases; these issues for locally linear actions on topological manifolds are still active areas of current research. The passage from algebraic information to geometric information at the level of CW-complexes in the non-free case was studied in [128] and [12].

The second missing ingredient in our discussion of the space form problem is the transition to manifolds. We have not yet addressed the existence question of when a Swan complex has the homotopy type of a manifold or the uniqueness question of classifying *all* manifolds within a homotopy type. Two aspects of these questions have been resolved nicely: the Madsen, Thomas and Wall result which gives the class of finite groups which arise as fundamental groups of manifolds covered by the sphere and Wall classification of fake lens spaces (i.e., manifolds homotopy equivalent to classical lens spaces) of dimension greater than five. Results on manifolds are usually accomplished via surgery theory and we discuss this next.

The systematic method for classification of manifolds is called *surgery*. The idea is that surgery theory reduces classification questions to a mix of algebraic topology and the algebra of quadratic forms. Some of the ingredients necessary for this reduction are handlebody theory, bundle theory, transversality, and embedding theorems. The embedding theorems make the theory most effective when the dimension of the manifold is $\geqslant 4$, where there is sufficient room to mimic algebraic manipulations by geometric embeddings. Both transversality and embedding theory provide "flies in the ointment" for the development of equivariant surgery. Hence, we concentrate on classical surgery theory.

As far as references for surgery theory, to the great detriment of the subject, there is no modern account. The most comprehensive is the book of Wall [166]. We refer the reader to the paper of Milnor [120] and the book of Browder [27] for geometric background. For background on classifying spaces and bundle theory see Milgram and Madsen [117]. For information specific to spherical space forms, see Davis and Milgram [52], Wall [166] and Madsen, Thomas and Wall [111]. Modern aspects of surgery theory can be found in the books of Ranicki [143] and Weinberger [172], however they were not written with classical surgery theory as their main focus.

3.3.4. *Existence of space forms.* When does a finite group G act freely on $\mathbb{S}^n$? As above, one must have $H^{n+1}(G, \mathbb{Z}) \cong \mathbb{Z}/|G|$. Then there exists a Swan complex, that is, a CW-complex X with dim $X = n$, $\pi_1(X) = G$, and $\widetilde{X} \simeq \mathbb{S}^n$. In fact, the k-invariant gives a one-to-one correspondence between polarized homotopy types and additive generators of $H^{n+1}(G, \mathbb{Z})$. Any Swan complex is finitely dominated, and the finiteness obstruction $[X] \in \widetilde{K}_0(\mathbb{Z}G)$ vanishes if and only if X has the homotopy type of a finite CW-complex. This can be effectively computed, [167,116,50], although the algebraic number theory can be quite involved. One qualitative result is the previously mentioned fact that if G is a group of period $n+1$, there is always a finite Swan complex of dimension $2n+1$.

To see when a finite Swan complex has the homotopy type of a manifold, one uses surgery theory, which provides necessary and sufficient conditions in dimensions greater than 4, and provides necessary conditions in all dimensions. First note that X is a *Poincaré complex*, i.e., X is a finite complex and there exists a class $[X] \in H_n(X)$ so that $\bigcap \mathrm{tr}[X]: H^i(X; \mathbb{Z}G) \to H_{n-i}(X; \mathbb{Z}G)$ is an isomorphism for all i.

The first obstruction to a Poincaré complex having the homotopy type of a manifold is the existence of a lift of the Spivak normal bundle to *BTOP*. The Spivak normal bundle is the homotopy theoretic analogue of the stable normal bundle of a manifold and is defined as follows. Embed X in $\mathbb{R}^K$ (K large) and let $N(X)$ be a closed regular neighborhood. Convert the map $\partial N(X) \to X$ to a fibration $p: E \to X$; then it is a formal consequence of Poincaré duality that the fiber has the homotopy type of $\mathbb{S}^{K-n-1}$ (see [27]). Fibrations with fibers having the homotopy type of a sphere are called *spherical fibrations*. Spherical fibrations over X are classified by $[X, BG]$ where $G = \operatorname{colimit} G_k$ is the stabilization of the topological monoid G_k of self-homotopy equivalences of $\mathbb{S}^{k-1}$. The map $p: E \to X$ and its classifying map $X \to BG$ are both referred to as the *Spivak normal bundle of* X. The next step is to see whether the Spivak bundle lifts to an honest topological sphere bundle, classified by a map $X \to BTOP$, and, if so, in how many ways. The obstruction to lifting the Spivak bundle to $BTOP$ is an element in $[X, B(G/TOP)]$. For a general Poincaré complex this might be non-trivial, but for a finite Swan complex it vanishes. The argument is as follows. The space $B(G/TOP)$ is an infinite loop space (i.e., the 0-space of an Ω-spectrum), and hence the abelian group $[X, B(G/TOP)]$ injects into $\bigoplus_p [X, B(G/TOP)]_{(p)}$ which injects into $\bigoplus_p [\widetilde{X}/G_p, B(G/TOP)]_{(p)}$, where G_p is a p-Sylow subgroup of G and the second injection is via the transfer map. Now $\widetilde{X}/G_p$ has the homotopy type of a lens space or a quaternionic space form, so the Spivak bundle map to $B(G/TOP)$ vanishes.

If the Spivak bundle $\nu: X \to BG$ of a Poincaré complex X lifts to $\tilde{\nu}: X \to BTOP$, then one can apply a transversality construction as a first step in the attempt to construct a manifold having the homotopy type of X.

DEFINITION 3.5. A *Thom invariant* for X is an element $\alpha \in \pi_{n+k}(T(\tilde{\nu}))$ for some lift $\tilde{\nu}: X \to BTOP(k)$ of the Spivak bundle ν, so that $h(\alpha) \cap U = [X]$ where h is the Hurewicz map and $U \in H^k(T(\tilde{\nu}))$ is the Thom class.

Given any lift $\tilde{\nu}$, Thom invariants always exist (see [27]) and are essentially given by collapsing out the complement of a regular neighborhood of X in $\mathbb{S}^{n+k}$. Given a Thom invariant $\alpha: \mathbb{S}^{n+k} \to T(\tilde{\nu})$, one may take the complement of the 0-section X to obtain a degree one map

$$f: M = \alpha^{-1}(X) \to X,$$

where M is a closed manifold. Furthermore, transversality gives a trivialization of $f^*(\tilde{\nu}) \oplus \tau_M$. Hence we call the induced map

$$(f, \hat{f}): (M, \nu_M) \to (X, \tilde{\nu})$$

a *degree one normal map*, where ν_M is the stable normal bundle of M (equipped with a trivialization of $\nu_M \oplus \tau_M$). For such a map, if $[\beta] \in \ker(\pi_i M \to \pi_i X)$ is represented by an embedding $\beta: \mathbb{S}^i \to X$ one may use the normal data $\hat{f}$ to thicken β up to an embedding $\mathbb{S}^i \times D^{n-i} \to M$ and perform surgery to obtain

$$M' = \left(M - \mathbb{S}^i \times \operatorname{int} D^{n-i}\right) \cup_{\mathbb{S}^i \times \mathbb{S}^{n-i-1}} \left(D^{i+1} \times \mathbb{S}^{n-i-1}\right)$$

and a degree one normal map $(g, \hat{g}): (M', \nu_{M'}) \to (X, \tilde{\nu})$. (In effect you are killing the homotopy class $[\beta]$, see [120].) The equivalence relation generated by surgery is called normal bordism and the set of equivalence classes is denoted $N(X)$. The main theorem of surgery theory is that there is the surgery obstruction map

$$\theta : N(X) \to L_n(\mathbb{Z}\pi_1 X).$$

$L_n(\mathbb{Z}\pi_1 X)$ is a algebraically defined abelian group associated to the group ring, closely related to quadratic forms over the group ring. If $(f, \hat{f})$ is normally bordant to a homotopy equivalence, then $\theta(f, \hat{f}) = 0$, and conversely, for $n > 4$, if $\theta(f, \hat{f}) = 0$, then $(f, \hat{f})$ is normally bordant to a homotopy equivalence.

An application to the spherical space form problem is given by:

THEOREM 3.11 (Petrie). *Let $G = \mathbb{Z}_7 \times_T \mathbb{Z}_3$ be the semidirect product where $\mathbb{Z}_3$ acts non-trivially on $\mathbb{Z}_7$. Then G acts freely on $\mathbb{S}^5$, but does not act freely and linearly.*

PROOF. It is an easy computation to show that G cannot act linearly and freely on $\mathbb{S}^5$. The Lyndon–Hochschild–Serre spectral sequence shows that $H^6(G, \mathbb{Z}) \cong \mathbb{Z}/21$, so G has period 6 and there is a Swan complex of dimension 5. Since $T(G) = \widetilde{K}_0(\mathbb{Z}G)$ (see [167, pp. 545–546]), there is a Swan complex with a zero finiteness obstruction, hence there is a finite Swan complex X of dimension 5 and fundamental group G. We have already observed that the Spivak bundle of X lifts to $BTOP$, so there is a degree one normal map

$$(f, \hat{f}) : (M, \nu_M) \to (X, \xi).$$

Now since G is odd order and n is odd, $L_n(\mathbb{Z}G) = 0$ (see [18]), one may complete surgery to a homotopy equivalence $M' \to X$. By the generalized Poincaré conjecture, $M' \cong \mathbb{S}^5/G$ for a free G-action of M'. (Note: a similar analysis gives a free smooth action on $\mathbb{S}^5$.) □

REMARK 3.1. Petrie [137] proved his theorem in a much more explicit and elementary manner. He noted that there is a free G-action on the Brieskorn variety

$$\Sigma^5 = \{(z_1, z_2, z_3, z_4) \in \mathbb{C}^4 \mid z_1^7 + z_2^7 + z_3^7 + z_4^3 = 0\} \cap \mathbb{S}^7.$$

This is only a rational homology sphere, but Petrie shows how to perform surgery to make it into a sphere. Another application of this sort of technique is given in [54].

The analysis of the problem of determining when a group G of period $n > 5$ can act freely on a sphere is much more difficult. One must compute surgery obstructions for all k-invariants and normal invariants. We state the beautiful result of Madsen, Thomas and Wall.

DEFINITION 3.6. Let n be a positive integer. A group G satisfies the *n-condition* if every subgroup of order n is cyclic.

THEOREM 3.12 (Madsen, Thomas and Wall [111]). *A finite group G acts freely on some sphere if and only if G satisfies the p^2- and $2p$-conditions for all primes p.*

In contrast, Wolf [174] analyzed the free, linear actions on spheres:

THEOREM 3.13. *A finite group G acts freely and linearly on some sphere if and only if the following two statements are satisfied.*

(i) *G satisfies the pq-condition for all primes p and q.*

(ii) *G has no subgroup isomorphic to $SL_2(\mathbb{F}_p)$ for a prime $p > 5$.*

3.3.5. *Uniqueness of space forms.* Surgery theory also attacks the uniqueness question – classifying manifolds within a homotopy type. It was motivated by the proof of the generalized Poincaré conjecture and the Kervaire–Milnor classification of exotic spheres. Two great successes of surgery theory are various cases of the Borel conjecture (especially that there are no fake tori of dimension greater than 3), and the classification of fake lens spaces due to Browder, Petrie, and Wall.

DEFINITION 3.7. A *fake lens space* is a space with covered by the sphere with cyclic fundamental group.

Every fake lens space is homotopy equivalent to a lens space.

THEOREM 3.14 ([166]). *Two fake lens spaces with odd order fundamental group and dimension greater than 3 are homeomorphic if and only if they have the same Reidemeister torsion and ρ-invariants.*

Furthermore, Wall shows exactly which invariants are realized. Reidemeister torsion was defined in Section 3.3.2. The ρ-invariant is defined as follows.

DEFINITION 3.8. For a closed, oriented, odd-dimensional manifold M with finite fundamental group G, define the *ρ-invariant*

$$\rho(M) = \frac{1}{s}\operatorname{sign}(G, W) \in \widetilde{K}_0(\mathbb{C}G) \otimes \mathbb{Q},$$

where s disjoint copies of the universal cover $\widetilde{M}$ bound a compact, oriented, free G-manifold W.

Given an action of a finite group G on a compact, oriented, even-dimensional manifold X, Atiyah and Singer [17, pp. 273–274], defined the G-signature $\operatorname{sign}(G, X) \in \widetilde{K}_0(\mathbb{C}G)$. We resist the urge to state the definition and to indicate the many applications in transformation groups of the Atiyah–Singer G-signature theorem.

3.4. *Final remarks*

In this paper we have attempted to cover the main ideas and examples which have made the subject of transformation groups an important and highly developed subject in topology. There are a number of important topics which we have not discussed here, such as actions of compact Lie groups, equivariant bordism, group actions on 4-manifolds, group actions on knot complements, the Nielsen Realization problem, etc. In the bibliography we have listed a number of references on these topics, so that hopefully the reader can access the original sources.

The background required to work on any remaining questions may require a combination of skills in manifold theory, group cohomology, algebraic K- and L-theory, homotopy theory and differential geometry. However daunting this may be, it is apparent that much remains to be understood about topological symmetries and we hope that the reader of this paper will take it upon himself to explore this topic further.

References

[1] A. Adem, *Cohomological restrictions on finite group actions*, J. Pure Appl. Algebra **54** (1988), 117–139.
[2] A. Adem, *Torsion in equivariant cohomology*, Comment. Math. Helv. **64** (1989), 401–411.
[3] A. Adem, *Automorphisms and cohomology of discrete groups*, J. Algebra **182** (1996), 721–737.
[4] A. Adem and W. Browder, *The free rank of symmetry of* $(\mathbb{S}^n)^k$, Invent. Math. **92** (1988), 431–440.
[5] A. Adem, R. Cohen and W. Dwyer, *Tate homology, homotopy fixed points and the transfer*, Contemporary Math. (1989).
[6] A. Adem and R.J. Milgram, *Cohomology of Finite Groups*, Grundlehren Math. Wiss., B. 309, Springer, Berlin (1994).
[7] A. Adem and J. Smith, *Periodic complexes and group actions*, Ann. of Math. (to appear).
[8] J.P. Alexander, P.E. Conner and G.C. Hamrick, *Odd Order Group Actions and Witt Classification of Inner-Products*, Lecture Notes in Math. 625, Springer, Berlin (1977).
[9] J.W. Alexander and G. Briggs, *On types of knotted curves*, Ann. of Math. **28** (1927), 562–586.
[10] C. Allday and V. Puppe, *Cohomological Methods in Transformation Groups*, Cambridge University Press (1993).
[11] M. Aschbacher and S. Smith, *On Quillen's conjecture for the p-groups complex*, Ann. Math. **137** (1993), 473–529.
[12] A.H. Assadi, *Finite group actions on simply-connected manifolds and CW-complexes*, Mem. Amer. Math. Soc. **35** (257) (1982).
[13] M.F. Atiyah, *K-theory*, Benjamin Pub. (1967).
[14] M.F. Atiyah and R.A. Bott, *A Lefschetz fixed point formula for elliptic complexes. II Applications*, Ann. of Math. **88** (1968), 451–491.
[15] M.F. Atiyah and R. Bott, *The moment map and equivariant cohomology*, Topology **23** (1984), 1–28.
[16] M.F. Atiyah, and G. Segal, *Equivariant K-theory and completion*, J. Differential Geom. **3** (1969), 1–18.
[17] M.F. Atiyah and I.M. Singer, *The index of elliptic operators. III*, Ann. of Math. **87** (1968), 546–604.
[18] A. Bak, *Odd dimensional surgery groups of odd torsion groups vanish*, Topology **14** (1975), 367–374.
[19] D. Benson, *Representations and Cohomology II*, Cambridge Studies in Advanced Math., Vol. 31, Cambridge Univ. Press (1991).
[20] D. Benson and J. Carlson, *Complexity and multiple complexes*, Math. Z. **195** (1987), 221–238.
[21] D. Benson and J. Carlson, *Projective resolutions and Poincaré duality complexes*, Trans. Amer. Math. Soc. **342** (1994), 447–488.
[22] R.H. Bing, *A homeomorphism between the 3-sphere and the sum of two solid horned spheres*, Ann. of Math. **56** (1952), 354–362.
[23] A. Borel, *Nouvelle Démonstration d'un Theoreme de P.A. Smith*, Comment. Math. Helv. **29** (1955), 27–39.
[24] A. Borel, *Seminar on Transformation Groups*, Princeton Univ. Press (1960).
[25] G. Bredon, *Introduction to Compact Transformation Groups*, Academic Press (1972).
[26] G. Bredon, *Equivariant Cohomology Theories*, Lecture Notes, Vol. 34, Springer, Berlin (1967).
[27] W. Browder, *Surgery on Simply-Connected Manifolds*, Springer, Berlin (1972).
[28] W. Browder, *Cohomology and group actions*, Invent. Math. **71** (1983), 599–608.
[29] W. Browder, *Actions of elementary abelian p-groups*, Topology **27** (1988), 459–472.

[30] W. Browder and W.C. Hsiang, *Some problems on homotopy theory manifolds and transformation groups*, Algebraic and Geometric Topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, CA, 1976), Part 2, Proc. Sympos. Pure Math. XXXII, Amer. Math. Soc., Providence, RI (1978), 251–267.
[31] W. Browder and W.C. Hsiang, *G-actions and the fundamental group*, Invent. Math. **65** (3) (1981/82), 411–424.
[32] W. Browder and F. Quinn, *A surgery theory for G-manifolds and stratified sets*, Manifolds, Tokyo 1973, Univ. of Tokyo Press (1975), 27–36.
[33] K. Brown, *Cohomology of Groups*, GTM, Vol. 87, Springer, Berlin (1982).
[34] S.E. Cappell and J.L. Shaneson, *Non-linear similarity*, Ann. of Math. **113** (1981), 315–355.
[35] S.E. Cappell, J.L. Shaneson, M. Steinberger and J.E. West, *Nonlinear similarity begins in dimension six*, Amer. J. Math. **111** (1989), 717–752.
[36] S.E. Cappell and S. Weinberger, *Homology propagation of group actions*, Comm. Pure Appl. Math. **40** (1987), 723–744.
[37] J. Carlson, *The varieties and cohomology ring of a module*, J. Algebra **85** (1983), 104–143.
[38] G. Carlsson, *On the rank of abelian groups acting freely on* $(\mathbb{S}^n)^k$, Invent. Math. **69** (1982), 393–400.
[39] G. Carlsson, *Equivariant stable homotopy theory and Segal's Burnside ring conjecture*, Ann. Math. **120** (1984), 189–224.
[40] G. Carlsson, *A survey of equivariant stable homotopy theory*, Topology **31** (1992), 1–27.
[41] H. Cartan and S. Eilenberg, *Homological Algebra*, Princeton University Press (1956).
[42] T. Chang and T. Skelbred, *Group actions on Poincaré duality spaces*, Bull. Amer. Math. Soc. **78** (1972), 1024–1026.
[43] D. Chase, *Fixed point sets of involutions on spheres*, Preprint.
[44] M.M. Cohen, *A Course in Simple-Homotopy Theory*, Springer, Berlin (1973).
[45] P.E. Conner, *Differentiable Periodic Maps*, Lecture Notes in Math., Vol. 738, Springer, Berlin (1979).
[46] P.E. Conner and E.E. Floyd, *Differentiable Periodic Maps*, Academic Press, New York; Springer, Berlin (1964).
[47] P.E. Conner and E.E. Floyd, *Maps of odd period*, Ann. of Math. (2) **82** (1966), 132–156.
[48] P.E. Conner, F. Raymond and P. Weinberger, *Manifolds with no periodic maps*, Proceedings of the Second Conference on Compact Transformation Groups, Part II, Lecture Notes in Math., Vol. 299, Springer, Berlin (1972), 81–108.
[49] M. Davis, W.C. Hsiang and J.W. Morgan, *Concordance classes of regular* $O(n)$*-actions on homotopy spheres*, Acta Math. **144** (3–4) (1980), 153–221.
[50] J.F. Davis, *Evaluation of the Swan finiteness obstruction*, Thesis, Stanford University (1982).
[51] J.F. Davis, *The homology of cyclic branched covers of* $\mathbb{S}^3$, Math. Ann. **301** (1995), 507–518.
[52] J.F. Davis and R.J. Milgram, *A Survey of the Spherical Space Form Problem*, Harwood (1984).
[53] J.F. Davis and R.J. Milgram, *Free actions on a product of spheres: The rational case*, Math. Z. **207** (1991), 379–390.
[54] J.F. Davis and T. tom Dieck, *Some exotic dihedral actions on spheres*, Indiana Univ. Math. J. **37** (2) (1988), 431–450.
[55] J.F. Davis and S. Weinberger, *Group actions on homology spheres*, Invent. Math. **86** (1986), 209–231.
[56] J.F. Davis and S. Weinberger, *Obstructions to propagation of group actions*, Bol. Soc. Mat. Mexicana **2** (1996), 1–14.
[57] M.W. Davis, *Groups generated by reflections and aspherical manifolds not covered by Euclidean space*, Ann. of Math. **117** (1983), 293–324.
[58] T. tom Dieck, *Actions of finite abelian p-groups without stationary points*, Topology **9** (1970), 359–366.
[59] T. tom Dieck, *Bordism of G-manifolds and integrality theorems*, Topology **9** (1970), 345–358.
[60] T. tom Dieck, *Characteristic numbers of G-manifolds, I*, Invent. Math. **13** (1971), 213–224.
[61] T. tom Dieck, *Equivariant homology and Mackey functors*, Math. Ann. **206** (1973), 67–78.
[62] T. tom Dieck, *Transformation Groups and Representation Theory*, Lecture Notes in Math., Vol. 766, Springer, Berlin (1979).
[63] T. tom Dieck, *Transformation Groups*, de Gruyter Studies in Math., Vol. 8 (1987).
[64] A. Dold and R. Thom, *Quasifaserungen und unendliche symmetrische Produckte*, Ann. of Math. **67** (1958), 239–281.
[65] M. Duflo and M. Vergne, *Cohomologie Equivariante et Descente*, Astérisque **215** (1993), 5–108.

[66] W. Dwyer, *Homology approximations to classifying spaces*, Topology **36** (1997), 783–804.
[67] W. Dwyer and C. Wilkerson, *Smith theory revisited*, Ann. Math. **127** (1988), 191–198.
[68] A.L. Edmonds and R. Lee, *Compact Lie groups which act on Euclidean space without fixed points*, Proc. Amer. Math. Soc. **55** (1976), 416–418.
[69] F.T. Farrell and W.C. Hsiang, *The topological-Euclidean space form problem*, Invent. Math. **45** (2) (1978), 181–192.
[70] F.T. Farrell and L.E. Jones, *Topological rigidity for compact non-positively curved manifolds*, Differential Geometry: Riemannian Geometry (Los Angeles, CA, 1990), Proc. Sympos. Pure Math., Vol. 54, Part 3, Amer. Math. Soc. (1993), 229–274.
[71] M.H. Freedman and F. Quinn, *Topology of 4-Manifolds*, Princeton Univ. Press (1990).
[72] D. Gabai, *On the geometric and topological rigidity of hyperbolic 3-manifolds*, J. Amer. Math. Soc. **10** (1997), 37–74.
[73] A.M. Gleason, *Spaces with a compact Lie group of transformations*, Proc. Amer. Math. Soc. **1** (1950), 35–43.
[74] D. Gottlieb, *The trace of an action and the degree of a map*, Trans. Amer. Math. Soc. **293** (1986), 381–410.
[75] J. Greenlees and P. May, *Generalized Tate cohomology*, Amer. Math. Soc. Mem. **113** (1995).
[76] I. Hambleton and E.K. Pedersen, *Non-linear similarity revisited*, Prospects in Topology (Princeton, NJ, 1994), Ann. of Math. Stud., Vol. 138, Princeton Univ. Press (1995), 157–174.
[77] R.S. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17** (1982), 255–306.
[78] F. Hirzebruch and T. Höfer, *On the Euler number of an orbifold*, Math. Ann. **286** (1990), 255–260.
[79] M. Hoffman, *Free actions of abelian groups on a Cartesian power of an even sphere*, Canad. Math. Bull. **30** (1987), 358–362.
[80] H. Hopf, *Zum Clifford–Kleinschen Raumproblem*, Math. Ann. **95** (1925), 313–319.
[81] W.Y. Hsiang, *On the classification of differentiable $SO(n)$ actions on simply connected π-manifolds*, Amer. J. Math. **88** (1966), 137–153.
[82] W.Y. Hsiang, *Cohomology Theory of Compact Transformation Groups*, Springer, Berlin (1975).
[83] W.C. Hsiang and W.Y. Hsiang, *Differentiable actions of compact connected classical groups I*, Ann. of Math. (2) **89** (1967), 705–786.
[84] W.C. Hsiang and W.Y. Hsiang, *Some problems in differentiable transformation groups*, Proc. Conf. on Transformation Groups (New Orleans, LA, 1967), Springer, New York (1967), 223–234.
[85] W.C. Hsiang and W.Y. Hsiang, *The degree of symmetry of homotopy spheres*, Ann. Math. **89** (1969), 52–67.
[86] W.C. Hsiang and W.Y. Hsiang, *Differentiable actions of compact connected classical groups II*, Ann. of Math. (2) **92** (1970), 189–223.
[87] W.C. Hsiang and W.Y. Hsiang, *Differentiable actions of compact connected Lie groups III*, Ann. of Math. (2) **99**, (1974), 220–256.
[88] W.C. Hsiang and W. Pardon, *When are topologically equivalent orthogonal transformations linearly equivalent?* Invent. Math. **68** (1982), 275–316.
[89] W.C. Hsiang and J.L. Shaneson, *Fake tori*, Topology of Manifolds (Proc. Inst., Univ. of Georgia, Athens, GA, 1969), Markham (1970), 18–51.
[90] S. Illman, *Smooth equivariant triangulations of G-manifolds for G a finite group*, Math. Ann. **233** (1978), 199–220.
[91] S. Illman, *The equivariant triangulation theorem for actions of compact Lie groups*, Math. Ann. **262** (1983), 487–501.
[92] S. Illman, *Smoothly equivalent real analytic proper G-manifolds are subanalytically equivalent*, Math. Ann. **306** (1996), 647–673.
[93] S. Illman, *Smooth proper G manifolds have unique equivariant triangulations and unique G-CW complex structures: Applications to equivariant Whitehead torsion*, Preprint.
[94] S. Jackowski and J. McClure, *Homotopy decompositions of classifying spaces via elementary abelian subgroups*, Topology **31** (1992), 113–132.
[95] S. Jackowski, J. McClure and R. Oliver, *Homotopy classification of self-maps of BG via G-actions I*, Ann. Math. **135** (1) (1992), 183–226.
[96] S. Jackowski, J. McClure and R. Oliver, *Homotopy classification of self-maps of BG via G-actions II*, Ann. of Math. **135** (2) (1992), 227–270.

[97] S. Jackowski, *Vector bundles over classifying spaces of compact Lie groups*, Acta Math. **176** (1) (1996), 109–143.
[98] L. Jeffrey and F. Kirwan, *Localization for non-abelian group actions*, Topology **34** (1995), 291–327.
[99] L. Jones, *A converse to the fixed-point theorem of P.A. Smith*, Ann. Math. **94** (1971), 52–68.
[100] K. Kawakubo, *The Theory of Transformation Groups*, Oxford Univ. Press (1991).
[101] S. Kerckhoff, *The Nielsen realization problem*, Ann. Math. **117** (1983), 235–265.
[102] R.C. Kirby and L.C. Siebenmann, *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations*, Ann. Math. Studies, Vol. 88, Princeton Univ. Press (1977).
[103] S. Kobayashi, *Transformation Groups in Differential Geometry*, Springer, Berlin (1972).
[104] N.H. Kuiper and J.W. Robbin, *Topological classification of linear endomorphisms*, Invent. Math. **19** (1973), 83–106.
[105] S. Kwasik and R. Schultz, *Icosahedral group actions on* $\mathbb{R}^3$, Invent. Math. **108** (1992), 385–402.
[106] E. Laitinen, M. Morimoto and K. Pawalowski, *Deleting-inserting theorem for smooth actions of finite non-solvable groups on spheres*, Comment. Math. Helv. **70** (1995), 10–38.
[107] R. Lee, *Semicharacteristic classes*, Topology **12** (1973), 183–199.
[108] W. Lück, *Transformation Groups and Algebraic K-theory*, Springer, Berlin (1989).
[109] I. Madsen, *Reidemeister torsion, surgery invariants, and spherical space forms*, Proc. London Math. Soc. (3) **46** (1983), 193–240.
[110] I. Madsen and M. Rothenberg, *On the classification of G-spheres. I. Equivariant transversality*, Acta Math. **160** (1988), 65–104.
[111] I. Madsen, C.B. Thomas and C.T.C. Wall, *The topological spherical space form problem II*, Topology **15** (1978), 375–382.
[112] A.I. Malcev, *On a class of homogeneous spaces*, Amer. Math. Soc. Transl. **39** (1951).
[113] G.A. Margulis, *Discrete Subgroups of Semisimple Lie Groups*, Springer, Berlin (1991).
[114] J. Martinet, *Modules sur l'algèbra du groupe quaternionien*, Ann. Sci. École Norm. Sup. **4** (1971), 299–308.
[115] J. Martino and S. Priddy, *The complete stable splitting for the classifying space of a finite group*, Topology **31** (1992), 143–156.
[116] R.J. Milgram, *Evaluating the Swan finiteness obstruction for manifolds for periodic groups*, Algebraic and Geometric Topology, Lecture Notes in Math., Vol. 1126, Springer, Berlin (1985), 127–158.
[117] R.J. Milgram and I. Madsen, *The Classifying Spaces for Surgery and Cobordism of Manifolds*, Princeton Univ. Press (1979).
[118] J. Milnor, *Construction of universal bundles I and II*, Ann. Math. **63** (1957).
[119] J. Milnor, *Groups which act on* $\mathbb{S}^n$ *without fixed points*, Amer. J. Math. **79** (1957), 623–630.
[120] J. Milnor, *A procedure for killing homotopy groups of differentiable manifolds*, Differential Geometry, Proc. Sympos. Pure Math., Vol. 3, Amer. Math. Soc., Providence, RI (1961), 39–55.
[121] J. Milnor, *A unique factorization theorem for three-manifolds*, Amer. J. Math. **84** (1962), 1–7.
[122] J. Milnor, *Whitehead torsion*, Bull. Amer. Math. Soc. **72** (1966), 358–426.
[123] J. Milnor, *Introduction to Algebraic K-Theory*, Princeton Univ. Press (1971).
[124] J.W. Morgan and H. Bass, eds, *The Smith Conjecture*, Academic Press, 1979.
[125] G.D. Mostow, *Factor spaces of solvable groups*, Ann. of Math. **60** (1954), 1–27.
[126] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Ann. of Math. Studies, Vol. 78, Princeton Univ. Press (1973).
[127] L.P. Neuwirth, *A topological classification of certain 3-manifolds*, Bull. Amer. Math. Soc. **69** (1963), 372–375.
[128] R. Oliver, *Fixed-point sets of group actions on finite acyclic complexes*, Comment. Math. Helv. **50** (1975), 155–177.
[129] R. Oliver, *A proof of the Conner conjecture*, Ann. Math. **103** (1976), 637–644.
[130] R. Oliver, *Smooth compact Lie group actions on disks*, Math. Z. **149** (1) (1976), 79–96.
[131] R. Oliver, *G-actions on disks and permutation representations*, J. Algebra **50** (1) (1978), 44–62.
[132] R. Oliver, *Free compact group actions on a product of spheres*, Algebraic Topology, Aarhus 1978, Lecture Notes in Math., Vol. 763, Springer, Berlin (1979), 539–548.
[133] R. Oliver, *Weight systems for SO(3)-actions*, Ann. of Math. (2) **110** (2) (1979), 227–241.

[134] R. Oliver, *A transfer for compact Lie group actions*, Proc. Amer. Math. Soc. **97** (3) (1986), 546–548.
[135] R. Oliver, *Fixed point sets and tangent bundles of actions on disks and Euclidean spaces*, Topology **35** (1996), 583–615.
[136] K. Pawalowski, *Fixed point sets of smooth group actions on disks and Euclidean spaces*, Topology **28** (1989), 273–289.
[137] T. Petrie, *Free metacyclic group actions on homotopy spheres*, Ann. of Math. **94** (1971), 108–124.
[138] T. Petrie, *One fixed point actions on spheres. I, II*, Adv. in Math. **46** (1982), 3–14, 15–70.
[139] T. Petrie and J.D. Randall, *Transformation Groups on Manifolds*, Marcel Dekker (1984).
[140] D. Quillen, *The spectrum of an equivariant cohomology ring I & II*, Annals Math. **94** (1971), 549–602.
[141] D. Quillen, *Homotopy properties of the poset of nontrivial p-subgroups of a group*, Adv. in Math. **28** (1978), 101–128.
[142] F. Quinn, *Ends of maps, II*, Invent. Math. **68** (1982), 353–424.
[143] A.A. Ranicki, *Algebraic L-Theory and Topological Manifolds*, Cambridge Univ. Press (1992).
[144] J.G. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Springer, Berlin (1994).
[145] U. Ray, *Free linear actions of finite groups on products of spheres*, J. Algebra **147** (1992), 456–490.
[146] F. Raymond, *Cohomological and dimension theoretical properties of orbit spaces of p-adic actions*, Proceedings of the Conference on Transformation Groups, New Orleans, 1967, Springer, Berlin (1968), 354–365.
[147] F. Raymond, *The role of Seifert fiber spaces in transformation groups*, Group Actions on Manifolds (Boulder, CO, 1983), Contemp. Math., Vol. 36, Amer. Math. Soc., Providence, RI (1985), 367–425.
[148] K. Reidemeister, *Knoten und Gruppen*, Abh. Math. Sem. Univ. Hamburg **5** (1927), 7–23.
[149] J.H. Rubinstein, *Free actions of some finite groups on* $\mathbb{S}^3$, Math. Ann. **240** (1979), 165–175.
[150] J. Rohlfs and J. Schwermer, *Special cycles*, J. Amer. Math. Soc. (1993).
[151] M. Rothenberg and S. Weinberger, *Group actions and equivariant Lipschitz analysis*, Bull. Amer. Math. Soc. **17** (1987), 109–112.
[152] C.U. Sanchez, *Actions of groups of odd order on compact, orientable manifolds*, Proc. Amer. Math. Soc. **54** (1976), 445–448.
[153] R. Schultz, *Nonlinear analogs of linear group actions on spheres*, Bull. Amer. Math. Soc. **11** (1984), 263–285.
[154] P. Scott, *There are no fake Seifert fibre spaces with infinite* π_1, Ann. of Math. **117** (1983), 35–70.
[155] G. Segal, *Equivariant K-theory*, Publ. Math. IHES **34** (1968), 129–151.
[156] J.L. Shaneson, *Linear algebra, topology and number theory*, Proceedings of the International Congress of Mathematicians, Vols 1, 2 (Warsaw, 1983), PWN (1984), 685–698.
[157] P.A. Smith, *Transformations of finite period*, Ann. of Math. **39** (1938), 127–164.
[158] P.A. Smith, *Transformation of finite period, II*, Ann. of Math. **40** (1927), 690–711.
[159] N. Steenrod, *The Topology of Fiber Bundles*, Princeton Univ. Press (1951).
[160] M. Steinberger, *The equivariant topological s-cobordism theorem*, Invent. Math. **91** (1988), 61–104.
[161] R.G. Swan, *Periodic resolutions for finite groups*, Ann. of Math. **72** (1960), 267–291.
[162] R.G. Swan, *A new method in fixed point theory*, Comment. Math. Helv. **34** (1960), 1–16.
[163] J. Thévenaz, *Alperin's conjecture*, J. Pure Appl. Algebra **85** (1993), 185–202.
[164] F. Waldhausen, *On irreducible 3-manifolds which are sufficiently large*, Ann. of Math. **87** (1968), 56–88.
[165] C.T.C. Wall, *Finiteness conditions for CW-complexes. II*, Proc. Roy. Soc. Ser. A **295** (1966), 129–139
[166] C.T.C. Wall, *Surgery on Compact Manifolds*, Academic Press (1970).
[167] C.T.C. Wall, *Periodic projective resolutions*, Proc. London Math. Soc. **39** (1979), 509–553.
[168] A. Wasserman, *Equivariant differential topology*, Topology **8** (1969), 127–150.
[169] P. Webb, *A local method in group cohomology*, Comment. Math. Helv. **62** (1987), 135–167.
[170] S. Weinberger, *Constructions of group actions: A survey of some recent developments*, Group Actions on Manifolds (Boulder, CO, 1983), Contemp. Math., Vol. 36, Amer. Math. Soc., Providence, RI (1985), 269–298.
[171] S. Weinberger, *Semifree locally linear PL actions on the sphere*, Israel J. Math. **66** (1989), 351–363.
[172] S. Weinberger, *The Topological Classification of Stratified Spaces*, Univ. of Chicago Press (1994).
[173] S. Weintraub *Semifree* $\mathbb{Z}/p$*-actions on highly connected manifolds*, Math. Z. **145** (1975), 163–185.
[174] J.A. Wolf, *Spaces of Constant Curvature*, McGraw-Hill (1967).

[175] E. Yalcin, *Group actions and group extensions*, Trans. Amer. Math. Soc. **352** (6) (2000), 2689–2700.
[176] C.T. Yang, *Hilbert's fifth problem and rel problems on transformation groups*, Mathematical Developments Arising from Hilbert Problems, Part. 1, Proc. Sympos. Pure Math., Vol. 27, Part 1, Amer. Math. Soc., Providence, RI (1976), 142–146.

CHAPTER 2

$\mathbb{R}$-Trees in Topology, Geometry, and Group Theory

Mladen Bestvina*

Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA
E-mail: bestvina@math.utah.edu

Contents

---

*Support by the National Science Foundation is gratefully acknowledged.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Abstract

An $\mathbb{R}$-tree is a metric space that can be represented as a monotonically increasing union of (metric) simplicial trees. Thus the set of vertices in an $\mathbb{R}$-tree may be dense. The importance of $\mathbb{R}$-trees stems from the fact that in many situations a sequence of negatively curved objects (manifolds, groups) gives rise ("converges") to an $\mathbb{R}$-tree together with a group acting on it by isometries. Their usefulness stems from the fact that this group can often be analyzed in terms of its point and arc stabilizers.

I will discuss several applications of the $\mathbb{R}$-tree technique, such as the finiteness of $Out(G)$ for many discrete groups G, compactness of certain representation spaces, and the analog of a 3-manifold decomposition theorem for arbitrary (finitely presented) groups.

1. Introduction

This paper is intended as a survey of the theory and applications of real trees from a topologist's point of view. The idea of an all-inclusive historical account was quickly abandoned at the start of this undertaking, but I hope to describe the main ideas in the subject with emphasis on applications outside the theory of $\mathbb{R}$-trees. The "Rips machine", i.e., the classification of measured laminations on 2-complexes, is the key ingredient. Roughly speaking, the Rips machine is an algorithm that takes as input a finite 2-complex equipped with a transversely measured lamination (more precisely, a band complex), and puts it in a "normal form". This normal form is surprisingly simple – the lamination is the disjoint union of finitely many sub-laminations each of which belongs to one of four types:

- *simplicial*: all leaves are compact and the lamination is a bundle over a leaf with compact 0-dimensional fiber,
- *surface*: geodesic lamination on a compact hyperbolic surface (or a cone-type orbifold),
- *toral*: start with a standard lamination of the n-torus by irrational planes of codimension 1 and restrict to the 2-skeleton; more generally, replace the torus by a cone-type orbifold covered by a torus (with the deck group leaving the lamination invariant),
- *thin*: this type is most interesting of all. It was discovered and studied by Levitt [29]. See Section 5.3 for the definition and basic properties.

Measured laminations on 2-complexes arise in the study of $\mathbb{R}$-trees via a process called *resolution*. In the simplicial case, this idea goes back to J. Stallings and was used with great success by M. Dunwoody. If G is a finitely presented group that acts by isometries on an $\mathbb{R}$-tree, one wants to deduce the structure of G, given the knowledge of vertex and arc stabilizers. Bass–Serre theory [55] solves this beautifully in the case of simplicial trees. For an exposition of Bass–Serre theory from a topological point of view, see [59].

I hope to convince the reader that the development of the theory of $\mathbb{R}$-trees is not an idle exercise in generalizations – indeed, in addition to the intrinsic beauty of the theory, $\mathbb{R}$-trees appear in "real life" as a brief look at the final section of this survey reveals. The reason for this is the construction presented in Section 3, which takes a sequence of isometric actions of G on "negatively curved spaces" and produces an isometric action of G on an $\mathbb{R}$-tree in the (Gromov–Hausdorff) limit.

The central part of the paper (Sections 4–6) is devoted to a study of the Rips machine and the structure theory of groups that act isometrically on $\mathbb{R}$-trees. The approach follows closely [5], and the reader is referred to that paper for more details. Gaboriau, Levitt, and Paulin have developed a different (but equivalent) point of view in a series of papers (see references, and in particular the survey [44] which puts everything together). For the historical developments and the state of the theory preceding Rips' breakthrough, see the surveys [56] and [57].

This paper is an expanded version of a talk presented at the AMS meeting #906 in Greensboro, NC in October 1995. I would like to thank Mark Feighn, Gilbert Levitt, and Zlil Sela for useful comments. Special thanks are due to G.A. Swarup for pointing out errors in the first version of this paper, and for correcting them. I will always be grateful to Mark Feighn for our long-term collaboration and for all the fun we had while learning and contributing to the mathematics described here.

2. Definition and first examples of $\mathbb{R}$-trees

DEFINITION 2.1. Let (X, d) be a metric space and let $x, y \in X$. An *arc from x to y* is the image of a topological embedding $\alpha : [a, b] \to X$ of a closed interval $[a, b]$ (and we allow the possibility $a = b$) such that $\alpha(a) = x$ and $\alpha(b) = y$. A *geodesic segment* from x to y is the image of an isometric embedding $\alpha : [a, b] \to X$ with $\alpha(a) = x$ and $\alpha(b) = y$.

DEFINITION 2.2. We say that (X, d) is an *$\mathbb{R}$-tree* if for any $x, y \in X$ there is a unique arc from x to y and this arc is a geodesic segment.

EXAMPLE 2.3. Let X be a connected 1-dimensional simplicial complex that contains no circles. For every edge e of X choose an embedding $e \to \mathbb{R}$. If $x, y \in X$, there is a unique arc A from x to y. This arc can be subdivided into subarcs $A_1, A_2, \dots, A_n$ each of which is contained in an edge of X. Define the length of A_i as the length of its image in $\mathbb{R}$ under the chosen embedding, and define $d(x, y)$ as the sum of the lengths of the A_i's. The metric space (X, d) is an $\mathbb{R}$-tree. We say that an $\mathbb{R}$-tree is *simplicial* if it arises in this fashion.

EXAMPLE 2.4 (*SNCF metric*). Take $X = \mathbb{R}^2$ and let e denote the Euclidean distance on X. Define a new distance d as follows. We imagine that there is a train line operating along each ray from the origin (= Paris). If two points $x, y \in X$ lie on the same ray, then $d(x, y) = e(x, y)$. In all other cases the train ride from x to y goes through the origin, so $d(x, y) = e(0, x) + e(0, y)$. The metric space (X, d) is a (simplicial) $\mathbb{R}$-tree.

EXAMPLE 2.5. A slight modification of the previous example yields a non-simplicial $\mathbb{R}$-tree. Take $X = \mathbb{R}^2$ and imagine trains operating on all vertical lines as well as along the x-axis. Thus $d(x, y) = e(x, y)$ when x, y are on the same vertical line, and $d(x, y) = |x_2| + |y_2| + |x_1 - y_1|$ otherwise, where we set $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

$\mathbb{R}$-trees that arise in applications tend to be separable (as metric spaces), and in fact they are the union of countably many lines. Example 2.5 can be easily modified to yield an example of a separable non-simplicial $\mathbb{R}$-tree (restrict to the subset of X consisting of the x-axis and the points with rational x-coordinate).

2.1. *Isometries of $\mathbb{R}$-trees*

I will now recall basic facts about isometric actions on $\mathbb{R}$-trees. Proofs are a straightforward generalization from the case of simplicial trees that can be found in [55]. Alternatively, the reader is referred to [37,13], or [1].

Let $\phi : T \to T$ be an isometry of an $\mathbb{R}$-tree T. The *translation length* of ϕ is the number

$$\ell(\phi) = \inf\{d(x, \phi(x)) \mid x \in T\},$$

where d denotes the metric on T. The infimum is always attained. If $\ell(\phi) > 0$ there is a unique ϕ-invariant line (= isometric image of $\mathbb{R}$), called the *axis* of ϕ, and the restriction

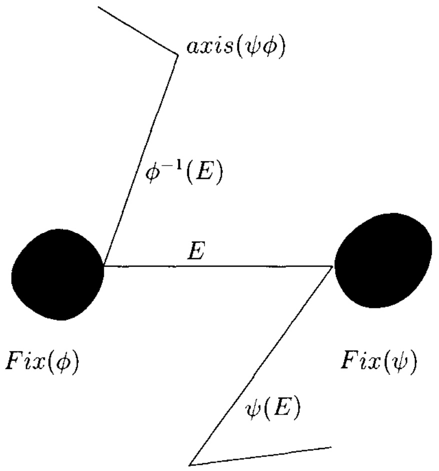


Fig. 1.

of ϕ to this line is translation by $\ell(\phi)$. In this case ϕ is said to be *hyperbolic*. If $\ell(\phi) = 0$, then ϕ fixes a non-empty subtree of T and is said to be *elliptic*.

EXERCISE 2.6. Let ϕ and ψ be two isometries of an $\mathbb{R}$-tree T. If they are both elliptic with disjoint fixed point sets, then the composition $\psi\phi$ is hyperbolic, and $\ell(\psi\phi)$ is equal to twice the distance between $Fix(\phi)$ and $Fix(\psi)$ (see Figure 1). If both ϕ and ψ are hyperbolic and their axes are disjoint, then $\psi\phi$ is hyperbolic, the translation length is equal to the sum of the translation lengths of ϕ and ψ plus twice the distance between the axes of ϕ and ψ, and the axis of $\psi\phi$ intersects both the axis of ϕ and of ψ.

EXERCISE 2.7. If $\{T_i\}_{i\in I}$ is a finite collection of subtrees of an $\mathbb{R}$-tree T such that all pairwise intersections are non-empty, then the intersection of the whole collection is non-empty.

Now let G be a group acting by isometries on an $\mathbb{R}$-tree T. The action is *non-trivial* if no point of T is fixed by the whole group. It is *minimal* if there is no proper G-invariant subtree.

EXERCISE 2.8. Use Exercise 2.7 to show that whenever a finitely generated group acts non-trivially on an $\mathbb{R}$-tree, then some elements of the group are mapped to hyperbolic isometries. Construct a (simplicial) counterexample to this statement when "finitely generated" is omitted from the hypotheses.

PROPOSITION 2.9. *Assume that G is finitely generated and that the action of G on T is non-trivial. Then T contains a unique G-invariant subtree $T' \subset T$ such that the action restricted to T' is minimal. Further, T' is the union of at most countably many lines.*

PROOF. Let T' be the union of the axes of hyperbolic elements in T. The only fact that needs a proof is that T' is non-empty and connected, and this follows from the exercises above. □

We will often replace a given $\mathbb{R}$-tree with the minimal subtree without saying so explicitly.

2.2. *δ-hyperbolic spaces – a review*

We recall the notion of "negative curvature" for metric spaces, due to Gromov [27]. For an inspired exposition see [23].

DEFINITION 2.10. Let (X, d) be a metric space and $* \in X$ a basepoint. For $x, y \in X$ define $(x \cdot y) = \frac{1}{2}(d(*, x) + d(*, y) - d(x, y))$. For $\delta \geqslant 0$ we say that $(X, *, d)$ is *δ-hyperbolic* if for all $x, y, z \in X$ we have

$$(x \cdot y) \geqslant \min\bigl((x \cdot z), (y \cdot z)\bigr) - \delta.$$

EXAMPLE 2.11. Let X be an $\mathbb{R}$-tree. Then $(x \cdot y)$ equals the distance between $*$ and the segment $[x, y]$. Further, if $x, y, z \in X$, then $(x \cdot y) \geqslant \min((x \cdot z), (y \cdot z))$. Thus $\mathbb{R}$-trees are 0-hyperbolic spaces. The converse is given in the next lemma.

EXAMPLE 2.12. Hyperbolic space $\mathbb{H}^n$ and any complete simply-connected Riemannian manifold with sectional curvature $\leqslant -\varepsilon < 0$ is δ-hyperbolic for some $\delta = \delta(\varepsilon)$.

If $(X, *, d)$ is δ-hyperbolic and if $*'$ is another basepoint, then $(X, *', d)$ is 2δ-hyperbolic. It therefore follows that the notions of "0-hyperbolic" and "hyperbolic" (i.e., δ-hyperbolic for some δ) don't depend on the choice of the basepoint.

A finitely generated group G is *word-hyperbolic* [27] if the word metric on G with respect to a finite generating set is hyperbolic. This notion is independent of the choice of the generating set.

The classification of isometries of hyperbolic spaces is more subtle than in the case of trees. Let $\phi : X \to X$ be an isometry of a hyperbolic metric space. The *translation length* can be defined as the limit

$$\ell(\phi) = \lim_{i \to \infty} \frac{1}{i} \inf_{x \in X} d\bigl(x, \phi^i(x)\bigr).$$

For a reasonable classification into hyperbolic, elliptic, and parabolic isometries it is necessary to assume something about X, e.g., that it is a geodesic metric space (any two points can be joined by a geodesic segment), or perhaps something weaker that guarantees that X does not have big holes. We will only need the following special case. If G is a word-hyperbolic group and $g \in G$ an element of infinite order, then left translation $t_g : G \to G$ by g has an *axis*, namely the set of points in G moved a distance $\leqslant \ell(t_g) + 10\delta$. ($G$ is δ-hyperbolic.) This set is quasi-isometric to the line.

2.3. *"Connecting the dots" lemma*

LEMMA 2.13. *Let $(X, *, d)$ be a 0-hyperbolic metric space. Then there exists an $\mathbb{R}$-tree (T, d_T) and an isometric embedding $i : X \to T$ such that*
(1) *no proper subtree of T contains $i(X)$, and*
(2) *if $j : X \to T'$ is an isometric embedding of X into an $\mathbb{R}$-tree T', then there is a unique isometric embedding $k : T \to T'$ such that $ki = j$.*

In particular, T is unique up to isometry. Further, if a group G acts by isometries on X, then the action extends to an isometric action on T.

PROOF. If $i : X \to T$ is an isometric embedding as in (1), then T is the union of segments of the form $I_x = [i(*), i(x)]$ for $x \in X$. The length of I_x is equal to $d(*, x)$, and two such segments I_x and I_y overlap in a segment of length $(x \cdot y)$. This suggest the construction of T. Start with the collection of segments $I_x = [0, d(*, x)]$ for $x \in X$ and then identify I_x and I_y along $[0, (x \cdot y)]$. For details, see, e.g., [40]. □

3. How do $\mathbb{R}$-trees arise?

Isometric actions of a group G on an $\mathbb{R}$-tree arise most often as the Gromov–Hausdorff limits of a sequence of isometric actions of G on a negatively curved space X. The construction is due independently to Paulin [41] and to Bestvina [2]. See also the expository article [10].

We will make the formal definition in terms of the projectivized space of equivariant pseudometrics on G.

3.1. *Convergence of based G-spaces*

Let G be a discrete group. By a *G-space* we mean a pair (X, ρ) where X is a metric space and $\rho : G \to Isom(X)$ is a homomorphism (an *action*) to the group of isometries of X. A *based G-space* is a triple $(X, *, \rho)$ where (X, ρ) is a G-space and $*$ is a basepoint in X that is not fixed by every element of G.

Recall that a *pseudometric* on G is a function $d : G \times G \to [0, \infty)$ that is symmetric, vanishes on the diagonal, and satisfies the triangle inequality. Let $\mathcal{D}$ denote the space of all pseudometrics ("distance functions") on G that are not identically 0, equipped with compact-open topology. We let G act on $G \times G$ diagonally, and on $[0, \infty)$ trivially, and consider the subspace $\mathcal{ED} \subset \mathcal{D}$ of G-equivariant pseudometrics. Scaling induces a free action of $\mathbb{R}^+$ on $\mathcal{ED}$, and we denote by $\mathcal{PED}$ the quotient space, i.e., the space of *projectivized equivariant distance functions* on G. A pseudometric on G is δ-hyperbolic if the associated metric space is δ-hyperbolic (the class of the identity element is taken to be the basepoint).

A based G-space $(X, *, \rho)$ induces an equivariant pseudometric $d = d_{(X,*,\rho)}$ on G by setting

$$d(g, h) = d_X\big(\rho(g)(*), \rho(h)(*)\big),$$

where d_X denotes the distance function in X. If the stabilizer under ρ of $*$ is trivial, then G can be identified with the orbit of $*$ via $g \leftrightarrow \rho(g)(*)$, and $d_{(X,*,\rho)}$ is the distance induced by d_X. We work with pseudometrics to allow for the possibility that distinct elements of G correspond to the same point of X.

DEFINITION 3.1. We say that a sequence $(X_i, *_i, \rho_i)$, $i = 1, 2, 3, \ldots$, of based G-spaces *converges* to the based G-space $(X, *, \rho)$ and write

$$\lim_{i \to \infty} (X_i, *_i, \rho_i) = (X, *, \rho)$$

provided $[d_{(X_i,*_i,\rho_i)}] \to [d_{(X,*,\rho)}]$ in $\mathcal{PED}$.

3.2. *Example: Flat tori*

To illustrate this, let us take $G = \mathbb{Z} \times \mathbb{Z}$ and $X = E^2$, the Euclidean plane. We will obtain actions of G on the real line as limits of discrete actions of $\mathbb{Z} \times \mathbb{Z}$ on E^2. The group $\mathbb{Z} \times \mathbb{Z}$ can act on E^2 by isometries in many different ways. We will only consider discrete orientation-preserving isometric actions, and those consist necessarily of translations and form the universal covering group of a flat 2-torus. Two such actions of $\mathbb{Z} \times \mathbb{Z}$ will be considered equivalent if there is a similarity of E^2 conjugating one action to the other, i.e., if the corresponding (marked) tori are conformally equivalent. It is convenient to identify the group of translations of E^2 with $\mathbb{C}$. Thus two actions $\rho_1, \rho_2 : G \to \mathbb{C}$ are equivalent if there is a complex number α such that $\rho_2(g) = \alpha\rho_1(g)$ for all $g \in \mathbb{Z} \times \mathbb{Z}$ or $\rho_2(g) = \overline{\alpha\rho_1(g)}$ for all $g \in \mathbb{Z} \times \mathbb{Z}$. Each equivalence class $[\rho]$ is uniquely determined by the complex-conjugate pair $\{z, \bar{z}\}$ where $z = \rho(0,1)/\rho(1,0)$, and thus the set of all equivalence classes can be identified with the upper half-plane $\{z \in \mathbb{C} \mid \operatorname{Im}(z) > 0\}$.

Let us take a sequence $\{z_n\}$ of points in the upper half-plane and let ρ_n be a representative of the equivalence class determined by z_n. Fix a basepoint $*$ in E^2.

PROPOSITION 3.2. *Suppose that $z_n \to r \in \mathbb{R} \cup \{\infty\}$. Then the sequence $(E^2, *, \rho_n)$ converges to the (unique up to scale) based G-space $(\mathbb{R}, 0, \rho)$ where the action ρ consists of translations and $\rho(0,1)/\rho(1,0) = r$.*

PROOF. Suppose for concreteness that $r \in \mathbb{R}$. We take ρ_n so that $\rho_n(1,0) = 1$ and $\rho_n(0,1) = z_n$. Then $\rho_n(g) \to \rho(g) \in \mathbb{R} \subset \mathbb{C}$ and the claim follows. □

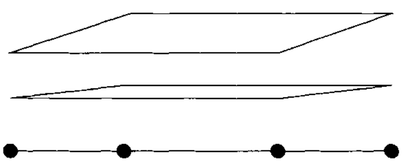


Fig. 2. The fundamental domain degenerates to a segment.

Thus this construction recovers the usual compactification of the upper half-plane by the circle $\mathbb{R} \cup \{\infty\}$.

If in this example we replace E^2 by E^n and $\mathbb{Z} \times \mathbb{Z}$ by $\mathbb{Z}^n$, the same construction would produce actions of $\mathbb{Z}^n$ by translations on E^m, $0 < m < n$, and would provide an equivariant compactification of the symmetric space $SL_n(\mathbb{R})/SO_n(\mathbb{R})$.

3.3. *$\mathbb{R}$-trees as limits of based δ-hyperbolic G-spaces*

The main reason for interest in $\mathbb{R}$-trees is the following result. Note that if a sequence of based G-spaces converges, the limit is far from being unique. In special situations we can take the limit to be an $\mathbb{R}$-tree.

THEOREM 3.3. *Let $(X_i, *_i, \rho_i)$ be a convergent sequence of based G-spaces. Assume that*

(1) *there exists $\delta \geqslant 0$ such that every X_i is δ-hyperbolic, and*

(2) *there exists $g \in G$ such that the sequence $d_{X_i}(*, \rho_i(g)(*))$ is unbounded.*

Then there is a based G-tree $(T, *)$ and an isometric action $\rho : G \to Isom(T)$ such that $(X_i, *_i, \rho_i) \to (T, *, \rho)$.

PROOF. The limiting pseudo-metric d on G is 0-hyperbolic, as it is the limit of pseudo-metrics $d_{(X_i, *_i, \rho_i)}/d_i$ with $d_i \to \infty$ (by (2)) and the ith pseudo-metric is δ/d_i-hyperbolic (by (1)). Now apply the connecting-the-dots Lemma 2.13 to the induced metric space. □

EXERCISE 3.4. Let $G = \mathbb{Z}$, $X_i = \mathbb{H}^2$ (hyperbolic plane), and the representation ρ_i sends the generator $1 \in \mathbb{Z}$ to a hyperbolic isometry whose translation length is 1 and whose axis passes at distance i from the basepoint in $\mathbb{H}^2$. Show that the limiting $\mathbb{R}$-tree T can be identified with the cone on $\mathbb{Z}$ and the limiting action ρ is the translation action on $\mathbb{Z}$ coned off. The basepoint is a point in $\mathbb{Z}$.

EXAMPLE 3.5. Let $f : F_n \to F_n$ be an automorphism of the free group $G = F_n = \langle x_1, \ldots, x_n \rangle$ of rank n that sends each basis element to a "positive word", i.e., a product of basis elements (not involving their inverses). Suppose that $\lambda > 0$ is the unique eigenvalue of the abelianization of f, viewed as an automorphism of $\mathbb{Z}^n$, with a corresponding eigenvector with non-negative coordinates $a_1, \ldots, a_n$. For X_i take F_n with the word metric, the basepoint is 1, and let ρ_i be the representation that sends $g \in F_n$ to the left translation by $f^i(g)$. The scaling factor can be taken to be $d_i = \lambda^i$. If $\lambda > 1$ we obtain in the limit an action ρ of F_n on an $\mathbb{R}$-tree. The positivity requirement was imposed to ensure that for some g the sequence of lengths of $f^i(g)$ grows at the "top speed", i.e., as (const)λ^i. The limiting tree can be described quite explicitly. For example, if g is a positive word, then the distance between the basepoint $* \in T$ and its image under $\rho(g)$ is $k_1 a_1 + \cdots + k_n a_n$, where k_j is the number of times the generator x_j appears in the word g.

More generally, this construction can be performed with the train-track maps of [6]. With the right choice of a train-track map one obtains free nonsimplicial isometric actions of the free group F_n ($n > 2$) on $\mathbb{R}$-trees (see [57]).

3.4. *Finding approximate subtrees*

We now assume that we are in the situation of Theorem 3.3 and we examine the limiting tree in more detail. Thus we assume that there is a sequence $d_i \to \infty$ such that

$$d_T\big(\rho(g)(*), \rho(h)(*)\big) = \lim_{i\to\infty} \frac{d_{X_i}(\rho_i(g)(*_i), \rho_i(h)(*_i))}{d_i}.$$

If x is a point in T that belongs to the orbit of $*$, then x can be "approximated" by the corresponding point $x_i \in X_i$ in the orbit of $*_i$. If two points in the orbit of $*$ coincide, then the corresponding points in X_i are "close" (more precisely, the distance between them divided by d_i goes to 0 as $i \to \infty$).

We now extend this discussion to all points of T. We will assume in addition that each X_i is a *geodesic metric space*, i.e., that any two points $x, y \in X_i$ are joined by a geodesic segment.

Let $x \in T$ be an arbitrary point. Fix a finite subset $F \subset G$ and define a set $X_i(F, x) \subset X_i$ "approximating" x to be the set of all points $x_i \in X_i$ that can be constructed as follows. Choose $g, h \in F$ so that x is on the segment in T connecting $\rho(g)(*)$ and $\rho(h)(*)$ and choose a geodesic segment in X_i connecting $\rho_i(g)(*_i)$ and $\rho_i(h)(*_i)$, and let x_i be the point on this segment that divides it in the same ratio as the point x divides the segment $[\rho(g)(*), \rho(h)(*)]$.

Of course, it might happen that $X_i(F, x) = \emptyset$ if there are no $g, h \in F$ as above. The following proposition summarizes the basic properties of this construction:

PROPOSITION 3.6.

(1) Equivariance: $X_i(gF, \rho(g)x) = \rho_i(g)X_i(F, x)$.

(2) Monotonicity: *If* $F \subset F'$ *then* $X_i(F, x) \subset X_i(F', x)$.

(3) Small diameter: $\frac{1}{d_i}$ diam $X_i(F, x) \to 0$ *as* $i \to \infty$.

(4) Metric convergence: *Let* $x, y \in T$. *Then for all finite* $F \subset G$ *and all choices* $x_i \in X_i(F, x)$, $y_i \in X_i(F, y)$ *we have*

$$\frac{1}{d_i} d_{X_i}(x_i, y_i) \to d_T(x, y).$$

(5) Non-triviality: *For every* $x \in T$ *there is a 2-element set* $F \subset G$ *such that* $X_i(F, x) \neq \emptyset$ *for all* i.

PROOF. Items (1) and (2) follow directly from the definition. Item (3) is an exercise in δ-hyperbolic geometry. Item (4) also follows directly from definitions if $F = \{g, h\}$ so that the segment in T joining $\rho(g)(*)$ and $\rho(h)(*)$ contains both x and y (the existence of such g, h follows from item (1) of Lemma 2.13, which also implies item (5)). The general case then follows from (2) and (3) by enlarging F. □

3.5. *Selecting the basepoint and the Compactness theorem*

We now assume that an action $\rho : G \to Isom(X)$ is given, and we consider the problem of locating a "most centrally located point" for this action. We will then use this point as the basepoint. In Example 3.4 the basepoint $*_i$ should be chosen on the axis of $\rho_i(1)$, and then the limiting action would be nontrivial.

The problem of finding a good basepoint has a satisfactory solution when G is finitely generated and X is a *proper* δ-hyperbolic metric space, and this is what we assume from now on. (A metric space is *proper* if closed metric balls are compact.) We also fix a finite generating set $S \subset G$. Let $F = F_{S,\rho} : X \to [0, \infty)$ be the function defined by

$$F(x) = \max_{g \in S} d_X\big(x, \rho(g)(x)\big).$$

The following lemma is an exercise.

LEMMA 3.7. *Assume that $\rho : G \to Isom(X)$ is non-elementary (i.e., it does not fix a point at infinity). Then $F : X \to [0, \infty)$ is a proper map. In particular, F attains its global minimum.*

We call a point $x \in X$ *centrally located (with respect to the action $\rho : G \to Isom(X)$ and the generating set S)* if F attains its global minimum at x.

PROPOSITION 3.8. *Suppose that under the hypotheses of Theorem 3.3 each X_i is proper and that the basepoints $*_i$ are centrally located (with respect to ρ_i and a fixed finite generating set S for G). Then the limiting action $\rho : G \to Isom(T)$ does not have global fixed points.*

PROOF. We can take $d_i = \max_{g \in S} d_{X_i}(*_i, \rho_i(g)(*_i))$. Suppose $x \in T$ is a global fixed point. Choose a finite subset $F \subset G$ so that $X_i(F, x) \neq \emptyset$. We will argue that for any $x_i \in X_i(F, x)$ and any $g \in S$ we have $\frac{1}{d_i} d_{X_i}(x_i, \rho_i(g)(x_i)) \to 0$ as $i \to \infty$, contradicting (for large i) the assumption that $*_i$ is centrally located. Indeed, $\rho(g)(x) = x$ coupled with equivariance property implies $X_i(gF, x) = \rho_i(g) X_i(F, x)$, so by monotonicity both x_i and $\rho_i(g)(x_i)$ belong to $X_i(gF \cup F, x)$, so the claim follows from the small diameter property. □

THEOREM 3.9 (Compactness theorem). *Suppose that (X, d) is a proper δ-hyperbolic metric space and that $\rho_i : G \to Isom(X)$ a sequence of non-elementary representations of a finitely generated group G. Assume that the group $Isom(X)$ acts cocompactly on X, i.e., that there is a compact subset $K \subset X$ whose $Isom(X)$-translates cover X. Then one of the following holds, possibly after passing to a subsequence.*

1. *There exist isometries $\phi_i \in Isom(X)$ such that the sequence of conjugates $\rho_i^{\phi_i}$ converges in the compact-open topology to a representation $\rho : G \to Isom(X)$.*
2. *For each i there exists a centrally located point $x_i \in X$ for the representation ρ_i such that the sequence of based G-spaces (X, x_i, ρ_i) converges to an action of G on an $\mathbb{R}$-tree T without global fixed points.*

PROOF. Let x_i be a centrally located point for ρ_i. If the sequence $d_i = \max_{g \in S} d_X(x_i, \rho_i(g)(x_i))$ converges to infinity, then item (2) holds, by the preceding proposition. Otherwise, after passing to a subsequence, the d_i's are uniformly bounded. In that case choose $\phi_i \in Isom(X)$ that sends x_i into $K \subset X$ and apply Arzela–Ascoli to the conjugates $\rho_i^{\phi_i}$ to see that item (1) holds in this case. □

In the situation (2) of the Compactness Theorem, assuming that X is a geodesic metric space or a hyperbolic group G equipped with a word metric, it can also be argued that the translation length $\ell(\rho(g))$ is equal to the limit $\lim_{i\to\infty} \ell(\rho_i(g))/d_i$.

3.6. *Arc stabilizers*

We now investigate, in the situation (2) of the Compactness Theorem, the arc stabilizers in the limiting action $\rho : G \to Isom(T)$. We restrict ourselves to two frequently encountered settings, when the arc stabilizers turn out to be "elementary".

Many reasonable groups, such as linear groups, satisfy the so called "Tits Alternative". This means that their subgroups are either "small" (virtually solvable) or "large" (contain a nonabelian free group). Word-hyperbolic groups satisfy a strong form of the Tits alternative: any subgroup either contains a nonabelian free group or it is virtually cyclic (elementary). Accordingly, an action of a group on an $\mathbb{R}$-tree is said to be *small* if it is non-trivial (there are no global fixed points), minimal, and all arc stabilizers are small.

PROPOSITION 3.10. *Let $H \subset G$ be the stabilizer under ρ of a non-degenerate arc in T.*

(1) *If each X_i is a copy of the Cayley graph Γ of a word hyperbolic group with respect to a fixed finite generating set, and each $\rho_i : G \to Isom(\Gamma)$ is a free action whose image is contained in the subgroup consisting of left translations, then H is virtually cyclic.*

(2) *If each X_i is a copy of a fixed rank* 1 *symmetric space* $\mathbb{H}$ *(real, complex, quaternionic hyperbolic space, or the Cayley plane), and each ρ_i is discrete and faithful, then H is virtually nilpotent.*

PROOF. Let $[a, b] \subset T$ be a non-degenerate segment fixed by H (under the action by ρ). Choose a sufficiently large finite subset $F \subset G$ and points $a_i \in X_i(F, a)$ and $b_i \in X_i(F, b)$. Let c_i be the midpoint on a geodesic segment σ_i connecting a_i and b_i.

(1) Say Γ is δ-hyperbolic. The key claim is that if $h, k \in H$ then, for large i, the left translation $\rho_i([h, k])$ moves c_i to a point at distance $< 20\delta$ from c_i. There is an upper bound to the number of left translations of Γ that move a given point a distance $\leqslant 20\delta$. Since the commutators $[h, k]$ for $h, k \in H$ generate the commutator subgroup $[H, H]$ of H, it follows from the freeness assumption that $[H, H]$ is finitely generated and in particular H is not a nonabelian free group. Since the same argument can be applied to any subgroup of H, we conclude that H does not contain a nonabelian free group, and hence it is virtually cyclic.

The idea of proof of the above key claim is that $\rho_i(h)$ and $\rho_i(k)$ map σ_i to a geodesic segment whose endpoints are within $\frac{1}{100}$ *length*(σ_i) of the endpoints of σ_i, and so these segments, except near the endpoints, run within 2δ of σ_i, i.e., $\rho_i(h)$ and $\rho_i(k)$ can be thought of (modulo small error) as translating along σ_i. Consequently, the commutator $\rho_i([h,k])$ fixes σ_i (modulo small error and away from the endpoints). Details are in [2] and [41].

(2) The proof here is a modification of (1), plus the Margulis lemma. Let μ be the Margulis constant for $\mathbb{H}$, so that if a discrete group of isometries of $\mathbb{H}$ is generated by isometries that move a point $x_0 \in \mathbb{H}$ a distance $< \mu$, then the group is virtually nilpotent. Arguing as in the key claim above, one can show that if $h, k \in H$, then for large i the isometry $\rho_i([h,k])$ moves c_i a distance $< \mu$. It then follows that every finitely generated subgroup of $[H, H]$ is virtually nilpotent and so $\rho_i(H)$ must be elementary (i.e., virtually nilpotent). □

3.7. *Stable actions*

DEFINITION 3.11. Suppose a group G is acting isometrically on an $\mathbb{R}$-tree T. A subtree of T is *non-degenerate* if it contains more than one point. A non-degenerate subtree $T_1 \subset T$ is said to be *stable* (with respect to the action) if for every non-degenerate subtree $T_2 \subset T_1$ we have the equality $Fix(T_1) = Fix(T_2)$ of pointwise stabilizers. The group action on T is *stable* if it is non-trivial, minimal, and every non-degenerate tree in T contains a stable subtree.

Group actions that tend to arise in practice are stable. For example, small actions of hyperbolic groups are stable. More generally, if the collection of arc stabilizers satisfies the ascending chain condition, then the (non-trivial and minimal) group action is stable. Note that if two stable subtrees of T have a non-degenerate intersection, then their union is a stable subtree. In particular, each stable subtree is contained in a unique maximal stable subtree.

The study of stable actions quickly reduces to the study of actions with trivial arc stabilizers (see Corollary 5.9 of [5]). To see the idea, assume that T is covered by maximal stable subtrees $\{T_i\}_{i\in I}$. Note that $T_i \cap T_j$ is at most a point for $i \neq j$. Now construct a simplicial tree S as follows. There are two kinds of vertices in S. There is a vertex for each maximal stable subtree T_i, and there is a vertex for each point of T that equals the intersection of distinct maximal stable subtrees. An edge is drawn from a vertex v of the first kind, determined by T_i, to the vertex w of the second kind, determined by $x \in T$, precisely when $x \in T_i$. The group G acts simplicially, without inversions of edges, on S. The stabilizer $Fix_S([v, w])$ of the edge $[v, w]$ described above fixes a point of T and the underlying assumption is that we understand arc and point stabilizers in T. We then appeal to Bass–Serre theory [55] to conclude that either G splits over an edge stabilizer in S or that G fixes a vertex of S. In the latter case, in view of nontriviality and minimality of the action of G on T, it follows that T itself is a stable tree, so after factoring out the kernel of the action, the induced action has trivial arc stabilizers.

4. Measured laminations on 2-complexes

We now review the basics of measured laminations. For more information and details the reader is referred to [38].

DEFINITION 4.1. A closed subset Λ of a locally path-connected metrizable space X is a *lamination* if every point $x \in \Lambda$ has a neighborhood U such that the pair $(U, U \cap \Lambda)$ is homeomorphic to the pair $(V \times (0,1), V \times C)$ for some topological space V and some compact totally disconnected subset $C \subset (0,1)$. Such a homeomorphism is called a *chart*. The path components of Λ are called *leaves*.

If X is a closed manifold, any codimension 1 (locally flat) submanifold is a lamination. More typically, the set C in the definition is the Cantor set.

EXAMPLE 4.2. Let X be a closed hyperbolic surface. Let γ_i be a sequence of simple closed geodesics in X. After possibly passing to a subsequence, this sequence converges in the Hausdorff metric to a closed subset Λ of X. One can check [11] that Λ is a lamination, and that the leaves of Λ are simple geodesics (closed or biinfinite). Such Λ is called a *geodesic lamination.*

DEFINITION 4.3. Let $\Lambda \subset X$ be a lamination and $\alpha : [a,b] \to X$ a path in X such that $\alpha(a), \alpha(b) \notin \Lambda$. We say that α is *transverse* to Λ if for every $t \in [a,b]$ with $\alpha(t) \in \Lambda$ there is a chart $h : (U, U \cap \Lambda) \to (V \times (0,1), V \times C)$ at $\alpha(t) \in X$ such that the map $pr_{(0,1)} h\alpha$ is a local homeomorphism at t.

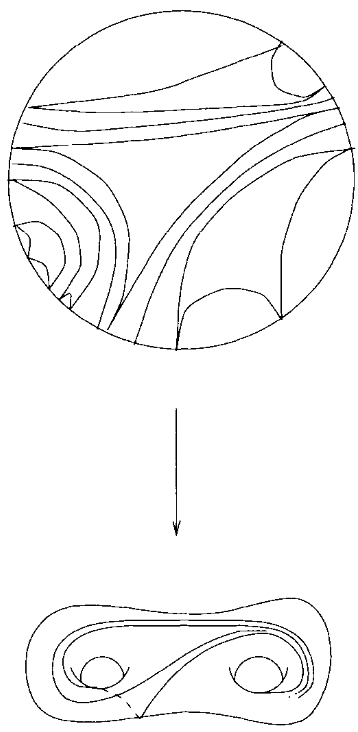


Fig. 3. A geodesic lamination on a hyperbolic surface and its universal cover. In the cover, countably many leaves are sides of ideal polygons. All other leaves are inaccessible from the complement.

DEFINITION 4.4. A *transverse measure* on a lamination $\Lambda \subset X$ is a function μ that assigns a nonnegative real number $\mu(\alpha)$ to every path α transverse to Λ and satisfies the following properties.

(1) If α is the concatenation of paths β and γ both of which are transverse to Λ, then $\mu(\alpha) = \mu(\beta) + \mu(\gamma)$.

(2) Every $x \in \Lambda$ has a chart $(U, U \cap \Lambda) \approx (V \times (0,1), V \times C)$ and there is a Borel measure ν on $(0,1)$ supported on C such that for any path $\alpha : [a,b] \to U$ with endpoints outside Λ that projects 1–1 to an interval in $(0,1)$, the measure $\mu(\alpha)$ equals the ν-measure of the projection.

The number $\mu(\alpha)$ is the *measure* of α.

EXERCISE 4.5. If two paths are homotopic through paths transverse to Λ, then they have the same measure. If a path is reparametrized, its measure does not change.

The *support* of μ is the complement of the set of points such that $\mu(\alpha) = 0$ whenever the image of α is contained in a sufficiently small neighborhood of the point. A lamination is *measured* if it is equipped with a transverse measure.

The support of μ is always a sublamination of Λ. We say that Λ has *full support* if the support is all of Λ.

EXAMPLE 4.6. Let Λ be a geodesic lamination on a hyperbolic surface X. If Λ is the finite union of simple closed curves, any transverse measure assigns a nonnegative real number, the *multiplicity* to each leaf, and the measure of any path transverse to Λ is the geometric intersection number with Λ, counted with multiplicity. Conversely, any such assignment determines a transverse measure. Now suppose that ℓ is an infinite leaf of Λ. We will construct a transverse measure on Λ, called the *counting* or the *hitting* measure. Triangulate the surface X so that the vertices are in the complement of Λ, all edges are geodesic segments, and each triangle is contained in a chart for Λ. For each edge e the intersection $e \cap \Lambda$ is totally disconnected. Choose a point in each component of $Int\, e \setminus \Lambda$. A transverse measure on Λ is determined by its values on the subintervals of the edges e with endpoints in the selected countable set. Conversely, if μ is defined on these countably many special intervals and the following two conditions hold, then μ extends uniquely to a transverse measure on Λ:

(1) (Additivity) If a special interval I is the concatenation of two special subintervals I_1 and I_2, then $\mu(I) = \mu(I_1) + \mu(I_2)$.

(2) (Compatibility) If I_1 and I_2 are special intervals belonging to two edges of the same triangle T in the triangulation, and if there is an embedded quadrilateral in T with two opposite sides I_1 and I_2, and the other two opposite sides disjoint from Λ, then $\mu(I_1) = \mu(I_2)$.

We will now construct a hitting measure on Λ. Choose a sequence of longer and longer closed subintervals $L_1, L_2, \ldots$ of the leaf ℓ. For a special interval I define

$$\mu_i(I) = \frac{N(L_i, I)}{N(L_i, X^{(1)})},$$

where $N(L_i, I)$ is the number of intersection points in $L_i \cap I$ and similarly $N_i = N(L_i, X^{(1)})$ is the number of intersection points between L_i and the 1-skeleton. Since ℓ is an infinite leaf, we have $N_i \to \infty$. Additivity and compatibility hold approximately for μ_i, i.e., in both cases the difference between the left-hand and the right-hand side is in the interval $[-\frac{1}{N_i}, \frac{1}{N_i}]$. Using a diagonalization process, pass to a subsequence if necessary so that $\lim \mu_i(I)$ exists for each special interval I, and set the limiting value equal to $\mu(I)$. The support of μ is generally smaller than Λ.

EXAMPLE 4.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear map that is injective when restricted to $\mathbb{Z}^n$ and consider the foliation of $\mathbb{R}^n$ by the level sets of f and the induced foliation $\mathcal{F}$ on the torus $T^n = \mathbb{R}^n/\mathbb{Z}^n$. The map f also defines a transverse measure on the two foliations. There is a standard way of converting the measured foliation $\mathcal{F}$ to a measured lamination Λ (and vice-versa); indeed, [5] is written in the language of foliations. More precisely, there is a map $T^n \to T^n$ whose point-preimages are arcs and points and the preimage of each leaf of $\mathcal{F}$ is either a leaf of Λ or the closure of a complementary component of Λ. This map is modeled on the Cantor function $[0, 1] \to [0, 1]$, which converts the foliation of $[0, 1]$ by points to the lamination on $[0, 1]$ whose underlying set is the Cantor set.

4.1. *Sacksteder's theorem*

We say that two paths γ and δ transverse to a lamination Λ are *pushing equivalent* if, after possibly reparametrizing one, they are homotopic through paths transverse to Λ. If Λ is equipped with a transverse measure μ, then by Exercise 4.5 we have $\mu(\gamma) = \mu(\delta)$. In particular, a measured lamination of full support satisfies the following *non-nesting condition*:

If $\gamma : [a, b] \to X$ and $\delta : [c, d] \to X$ are pushing-equivalent and γ is a subpath of δ (i.e., $\gamma = \delta|[a, b]$), then $\delta([c, d] \setminus [a, b]) \subset X \setminus \Lambda$.

There is a remarkable converse, due to R. Sacksteder.

THEOREM 4.8 [49]. *Suppose X is compact and $\Lambda \subset X$ is a lamination on X satisfying the above non-nesting condition. Then there is a non-trivial transverse measure on Λ, possibly not of full support.*

It is easy to construct examples of non-nesting laminations on compact spaces that do not support a transverse measure. For example, take a geodesic measured lamination on a hyperbolic surface and replace a noncompact leaf with a parallel family of leaves. For an $\mathbb{R}$-tree version of Sacksteder's theorem, see [30].

4.2. *Decomposition into minimal and simplicial components*

We say that a lamination $\Lambda \subset X$ is *simplicial* if there is a leaf ℓ of Λ, a closed neighborhood N of Λ in X and a map $N \to \ell$ which is an I-bundle and whose restriction to Λ is a bundle map with 0-dimensional fibers. A lamination Λ is *minimal* if every leaf of Λ is dense in Λ. When the underlying space X is compact, the lamination that supports a transverse measure always decomposes into simplicial and minimal sub-laminations.

THEOREM 4.9 (Theorem 3.2 in [38]). *Let X be compact and $\Lambda \subset X$ a lamination that admits a transverse measure with full support. Then Λ is the disjoint union $\Lambda_1 \sqcup \Lambda_2 \sqcup \cdots \sqcup \Lambda_n$ with each Λ_i either simplicial or minimal.*

On a closed hyperbolic surface, imagine a lamination consisting of two closed geodesics and a biinfinite geodesic that spirals towards the closed geodesics, one in each direction. Such a lamination does not decompose into simplicial and minimal sub-laminations. It is also not hard to show directly that this lamination does not support a transverse measure; indeed, this lamination is not even non-nesting.

4.3. *Resolutions*

Let G be a finitely presented group, and assume that G is acting non-trivially and minimally on an $\mathbb{R}$-tree T (as usual, by isometries). Since G is finitely presented, there is a finite simplicial complex K of dimension $\leqslant 2$ whose fundamental group is G. We now use T to construct a measured lamination Λ on K and an equivariant map $f : \widetilde{K} \to T$ from the universal cover of K to T that sends leaves of the preimage lamination $\tilde{\Lambda} \subset \widetilde{K}$ to points. We refer to this map as a *resolution*. In the case of simplicial trees this construction has been extensively used by M. Dunwoody (the leaves in this case are Dunwoody's "tracks").

To construct Λ and f, first choose a countable equivariant dense subset $D \subset T$ that includes all branch points of T ($v \in T$ is a branch point if the tripod, i.e., the cone on 3 points, can be embedded in T with the cone point mapped to v) and that intersects each arc in a dense set. This is possible by Proposition 2.9. Then define f on the vertices of $\widetilde{K}$ so that the map is equivariant and sends each vertex into D. Next, extend f equivariantly to the edges of $\widetilde{K}$. If the endpoints of a given edge e map to the same point under f, then define f on e to be the constant map. Otherwise, $f|e$ is chosen so that it is the Cantor function onto the arc whose boundary is $f(\partial e)$ with the preimage of each point in $D \cap f(e)$ an arc and the preimage of every other point in $f(e)$ a single point. The Cantor set of points in e that don't belong to the interior of a preimage arc is going to be the set $\tilde{\Lambda} \cap e$. Finally, extend f equivariantly to each 2-simplex σ of K so that for each $y \in D \cap f(\sigma)$ the preimage $f^{-1}(y) \cap \sigma$ is a convex triangle, quadrilateral, or a hexagon with vertices in $\partial\sigma$ and the preimage of every other point in $f(\sigma)$ is a straight line segment joining two distinct sides of σ.

These line segments are the components of $\tilde{\Lambda} \cap \sigma$. The transverse measure is defined by the requirement that if α is a path in $\widetilde{K}$ that is transverse to $\tilde{\Lambda}$ and intersects each leaf at most once, then the measure of α is the distance in T between the f-images of the endpoints of α. This transverse measure is equivariant and descends to a transverse measure on the induced lamination $\Lambda \subset K$.

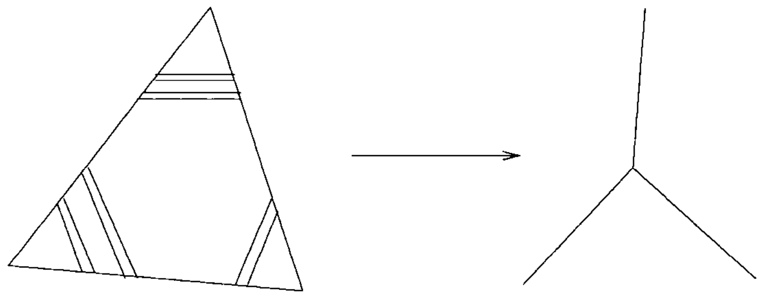


Fig. 4.

4.4. *Dual trees*

There is a construction that to a measured lamination Λ on a finite complex K assigns an $\mathbb{R}$-tree on which the fundamental group of the complex acts. Let $\widetilde{K}$ be the universal cover of K and $\tilde{\Lambda}$ the induced lamination on $\widetilde{K}$. Define a pseudometric $d : \widetilde{K} \times \widetilde{K} \to [0, \infty)$ by taking

$$d(x, y) = \inf_{\alpha} \tilde{\mu}(\alpha),$$

where $\tilde{\mu}$ is the induced transverse measure and the infimum runs over all paths that are transverse to $\tilde{\Lambda}$ and join x to y. It is not difficult to show that the associated metric space T is an $\mathbb{R}$-tree, called the *dual tree*, and that the deck group induces an isometric action of G on T. There is also the natural quotient map $f : \widetilde{K} \to T$; it is equivariant and maps each leaf and each complementary component of $\tilde{\Lambda}$ to a point.

In general, many different leaves will map to the same point by f. For example, start with a geodesic lamination on a 4 times punctured sphere and then fill in the punctures. The dual tree in this case is a single point. It is reasonable to impose the condition that f restricted to each edge e of $\widetilde{K}$ is the Cantor function that collapses precisely the closures of complementary components of $e \cap \Lambda$ in e. This condition is automatically satisfied when Λ arises as in the construction of a resolution. For the lack of a better term, we say that f is *locally injective* if it satisfies this condition.

QUESTION. Assume that f is locally injective. Is the infimum above always realized by a "minimizing" path α? Can f map distinct leaves of $\tilde{\Lambda}$ that do not belong to the closure of the same complementary component to the same point?

EXAMPLE 4.10. Let Λ be a geodesic measured lamination on a closed hyperbolic surface such that the measure has full support and such that the complementary components are simply-connected (see, e.g., [11]). The action of the fundamental group of the surface on the dual tree is free.

Notice that the constructions of a resolution and of the dual tree are generally not inverses of each other. For example, a free group admits many interesting non-simplicial actions on $\mathbb{R}$-trees (e.g., via the construction as in the preceding example applied to a punctured surface), while the dual of any resolution that uses a bouquet of circles for K is simplicial.

A resolution $f:\widetilde{K} \to T$ is *exact* (see [5]) if all point preimages are connected. This is equivalent to the statement that each point preimage is either a leaf or the closure of a complementary component. A group action on an $\mathbb{R}$-tree T is *geometric* if it admits an exact resolution. For more information on geometric actions, see [31].

Frequently, one encounters the following situation: $f:\widetilde{K} \to T$ is a resolution, $\tilde{\Lambda}$ the associated lamination, and $f':\widetilde{K} \to T'$ is the equivariant map to the tree dual to $\tilde{\Lambda}$. By construction, we have a factorization

$$f = \pi f'$$

for an equivariant map $\pi : T' \to T$. As remarked above, this map may not be an isometry. If f is an exact resolution, then π is an isometry.

QUESTION. If π is an isometry, is f an exact resolution?

It is a consequence of the Rips machine that if the action on T is stable, then $f':\widetilde{K} \to T'$ is an exact resolution, so the potential pathologies in the questions above don't arise in the stable case.

In general, one can say that if f is not exact, then either π is not an isometry, or there exist two leaves of $\tilde{\Lambda}$ that can be joined by a path with arbitrarily small measure, but cannot be joined by a path of measure 0. In either case, there are two leaves of $\tilde{\Lambda}$ such that any path joining them has measure strictly larger than the distance between their f-images x and y. One can then construct a "better resolution" as follows. Choose a path in $\widetilde{K}$ joining two such leaves. Attach a 2-cell $[0,1] \times [0,1]$ by gluing $[0,1] \times 0$ to the path. Map the other 3 boundary components to the arc $[x,y] \subset T$ (point if $x = y$). Then extend f to the 2-cell in the same way as when constructing a resolution. Finish the construction by attaching the whole orbit of 2-cells and extending to preserve equivariance. Slight care and subdivisions may be necessary to stay in the simplicial category. In the end, we have another resolution $f':\widetilde{K}' \to T$ and a factorization $f = f'\rho$, where $\rho:\widetilde{K} \to \widetilde{K}'$ is equivariant, sends leaves to leaves, and the images in $\widetilde{K}'$ of the original pair of leaves are joined by a path whose measure is equal to the distance between x and y.

Continuing in this fashion, we can construct resolutions that more and more faithfully reflect the nature of T.

PROPOSITION 4.11. *Assume that a finitely presented group G is acting by isometries on an $\mathbb{R}$-tree T and the action is non-trivial and minimal. For any finite collection $Y \subset T$ of points in T and any finite collection $G_0 \subset G$ of group elements there is a resolution $f:\widetilde{K} \to T$ (K depends on Y and G_0) and a collection of points $Y' \subset \widetilde{K}$ such that $\tilde{f}$ induces a bijection between Y' and Y and for any $a, b \in Y'$ and any $\gamma, \delta \in G_0$ there is a path α from $\gamma(a)$ to $\delta(b)$ whose measure is equal to $d(f(\gamma(a)), f(\delta(b)))$.*

For example, if $x \in T$ and H is a finitely generated subgroup of the stabilizer $Stab(x)$, then we can construct a resolution $f : \widetilde{K} \to T$ such that f sends a leaf or a complementary component D to x and $h(D) = D$ for all $h \in H$.

On the other hand, using the construction outlined in Example 3.5, one can show that there are examples of free actions of the free group F_3 such that every resolution is simplicial [3].

4.5. *Band complexes*

It is more convenient to work with a special class of 2-complexes equipped with with measured laminations, called *band complexes.*

DEFINITION 4.12. A *band* is the square $[0, 1] \times [0, 1]$ equipped with a measured lamination $C \times [0, 1]$ with measure of full support for a compact totally disconnected set $C \subset (0, 1)$.

A *multiinterval* Γ is the disjoint union of closed intervals equipped with a measured lamination $\Lambda(\Gamma)$ disjoint from the endpoints.

A *union of bands* is the space Y obtained from a multiinterval Γ by attaching a collection of bands. Each band $[0, 1] \times [0, 1]$ is attached via an embedding $\phi : [0, 1] \times \{0, 1\} \to \Gamma$ such that $\phi^{-1}(\Lambda(\Gamma)) = C \times \{0, 1\}$ and such that ϕ is measure-preserving. The measured lamination $\Lambda(\Gamma)$ pieces together with the measured laminations on the bands to produce a measured lamination $\Lambda(Y)$ on Y.

A *band complex* is the space X obtained from a union of bands Y by successively attaching 0-, 1-, and 2-cells (with PL attaching maps) so that

- There is a neighborhood of $\Lambda(Y)$ disjoint from the images of all attaching maps.
- The images of attaching maps of 1-cells are contained in $\Gamma \cup$ 0-cells.

The band complex X is equipped with the induced measured lamination $\Lambda = \Lambda(X)$.

EXAMPLE 4.13. Let X be the hyperbolic surface of Example 4.6. Each triangle in X intersects the lamination Λ in a collection of geodesic arcs, each spanning between two sides. Thus these arcs fall into at most 3 families according to which two sides they intersect. We can view X as a band complex as follows. The multi-interval Γ is obtained from the 1-skeleton by removing small disks around each vertex. Each triangle gives rise to at most 3 bands, one for each family of geodesic arcs. The vertices are the 0-cells, there are two 1-cells for each edge of the triangulation, connecting an endpoint to Γ. Finally, a triangle of the most interesting type (intersecting Λ in 3 families of arcs) gives rise to four 2-cells, three corner triangles, and a central hexagon. Simpler triangles give rise to fewer 2-cells.

DEFINITION 4.14. Let X be a band complex and assume that $\pi_1(X)$ is acting on an $\mathbb{R}$-tree T. An equivariant map $f : \widetilde{X} \to T$ is a *resolution* (or an *exact resolution*) if there is a triangulation of X so that f is a resolution (or an exact resolution) in the sense of Section 4.3.

5. Rips machine

5.1. *Moves on band complexes*

Building on the work of Makanin [34] and Razborov [45], Rips has devised a "machine" that transforms any band complex into a "normal form". The reference for this section is [5]. Here we only outline some aspects of the Rips machine.

There is a list of 6 moves M0–M5 that can be applied to a band complex. The complete list is in Section 6 of [5]. These moves are analogs of the elementary moves in simple homotopy theory, but they respect the underlying measured lamination. If a band complex X' is obtained from a band complex X by a sequence of these moves, then the following holds.

- There are maps $\phi : X \to X'$ and $\psi : X' \to X$ that induce an isomorphism between fundamental groups and preserve measure.
- If $f : \widetilde{X} \to T$ is a resolution, then the composition $f\tilde{\psi} : \widetilde{X}' \to T$ is also a resolution, and if $g : \widetilde{X}' \to T$ is a resolution, then so is $g\tilde{\phi} : X \to T$.
- ϕ and ψ induce a 1–1 correspondence between the minimal components of the laminations on X and X'.
- $\tilde{\phi}$ and $\tilde{\psi}$ induce quasi-isometries between the leaves of the laminations in $\widetilde{X}$ and $\widetilde{X}'$.

By way of illustration, we describe one of the moves, namely (M5). An arc $J \subset \Gamma$ is said to be *free* if the endpoints of J are in the complement of Λ, J has positive measure, and it intersects only one attaching region of a band. A free subarc J is said to be a *maximal* free subarc if whenever $J' \supset J$ is a free subarc, then $J' \cap \Lambda = J \cap \Lambda$.

Assume that J is a maximal free subarc and that J is contained in the attaching region $[0, 1] \times 0$ of a band $B = [0, 1] \times [0, 1]$. The move (M5) consists of collapsing $J \times [0, 1]$ to $J \times 1 \cup Fr\ J \times [0, 1]$. Typically, the band B will be replaced by two new bands, but if J contains one or both endpoints of the attaching region $[0, 1]$, then B is replaced by 1 or 0 bands. Attaching maps of relative 1- and 2-cells whose images intersect $int\, J \times [0, 1)$, can be naturally homotoped upwards.

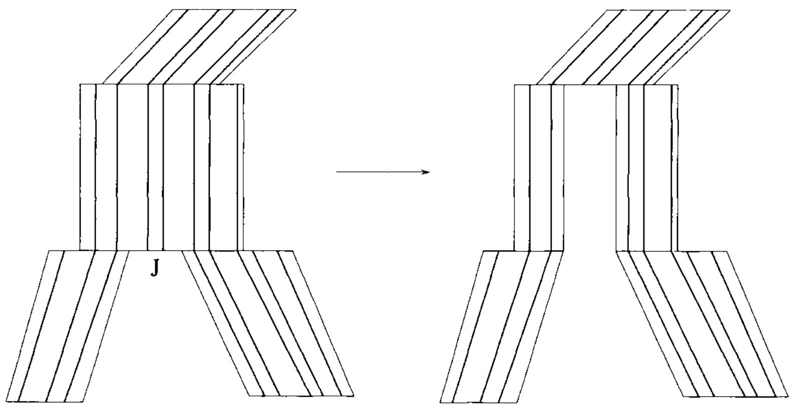


Fig. 5.

5.2. *The classification theorem*

For simplicity, we will assume that X is a band complex and $f : \widetilde{X} \to T$ is a resolution of an $\mathbb{R}$-tree T on which $G = \pi_1(X)$ is acting, and that the action has trivial arc stabilizers. This assumption is not necessary, but it dramatically simplifies the statements. The reason is that in this case it is always possible to remove annuli from a band complex. Imagine a band complex X that contains as a subcomplex an annulus $[0, 1] \times S^1$ which is thought of as a single band with top and bottom attached to the same arc. If the measure of the arc $[0, 1] \times p$ is positive, then the element of the fundamental group corresponding to the loop $q \times S^1$ fixes an arc in T and is therefore trivial. We can then collapse the annulus to the arc and replace X by the resulting complex X' (this is move (M1)).

We will also assume that $\pi_1(X)$ is torsion-free.

THEOREM 5.1 (Rips [5]). *Let X be a band complex such that $\widetilde{X}$ resolves an action of the torsion-free group $\pi_1(X)$ on an $\mathbb{R}$-tree with trivial arc stabilizers. Then X can be transformed, using moves* (M0–M5), *to another band complex X' with the following properties. For each minimal component Λ'_i of the lamination Λ' on X' there is a subcomplex X'_i of X' that intersects Λ' in Λ'_i and these subcomplexes are pairwise disjoint. All inclusions $X'_i \hookrightarrow X'$ are π_1-injective, and so are all inclusions from a component of the frontier $\dot{Fr}(X'_i)$ into X'_i. Each X'_i is of one of the following* 3 *types*:

- Surface type: *X'_i is a compact surface with negative Euler characteristic and Λ'_i is a (filling) geodesic measured lamination (with respect to a hyperbolic structure on the surface). Each component of $Fr(X'_i)$ is either a point or a boundary component of the surface.*
- Toral type: *X'_i is the* 2*-skeleton of the torus from Example* 4.7 *with the induced lamination. Each component of $Fr(X'_i)$ is a point.*
- Thin type: *This type does not have a standard model. Its main feature is that it can be arranged that X'_i contains an arbitrarily thin band (i.e., with attaching regions of small measure) that intersects the rest of X' only in the two attaching regions. See more on this below.*

If $\pi_1(X)$ is not torsion-free, the theorem still holds provided that in the surface and toral types we allow for a finite number of cone-type orbifold points.

Traditionally, in terms of the dual tree (or the associated pseudogroup), surface type is called "interval exchange", and toral type is called "axial". Similarly, thin type is also called "Levitt type", in honor of Levitt [29] who discovered and extensively studied this kind of a pseudogroup. Thin type has also been called "exotic". We chose names that reflect the nature of the band complex, not the dual tree or the pseudogroup.

5.3. *Thin type band complexes*

We now describe band complexes of thin type in more detail. Suppose X is a band complex whose lamination Λ is minimal. If there are maximal free arcs, choose one and perform the collapsing move (M5) described above to obtain another band complex X_1. If X_1 contains maximal free arcs, choose one and collapse, etc. This process stops when there are no more free arcs. Note that a collapse might produce new free arcs. This process is called Process I in [5].

DEFINITION 5.2. The band complex X is of *thin type* if it is equivalent under moves (M0–M5) to a complex for which the collapsing procedure never ends.

There is a concrete example of a thin band complexes in Section 10 of [5]. This example has the additional feature that after each collapse the resulting band complex is a scaled down version of the original. Martin [35] has studied the relation between the "periodicity" of the sequence of collapses and unique ergodicity of the underlying lamination. The band complexes associated to the "interesting" pseudogroups in [29] are thin.

If we focus on a single band in X, then under the collapsing process this band will get subdivided into more and more bands with arbitrarily small transverse measure (so these bands are thin, thus the name). In particular, eventually there will be bands whose interiors are disjoint from the attaching regions of the relative 2-cells. Such bands are called *naked bands*. A naked band induces a free product decomposition of $\pi_1(X)$ by cutting along an arc in the band that separates the two attaching regions. That this decomposition is non-trivial is the content of Proposition 8.13 of [5].

As the reader will learn from Section 7, in applications one frequently assumes that the underlying group $\pi_1(X)$ is freely indecomposable, and then the situation simplifies considerably as there can be no thin components in resolving band complexes (of course, assuming the arc stabilizers of T are trivial). Similarly, when one is concerned with hyperbolic groups, there can be no toral components.

One can also make a study of quasi-isometry types of leaves for a thin type lamination. Generic leaves are quasi-isometric to 1-ended trees, and in addition there are uncountably many leaves quasi-isometric to 2-ended trees. For details see Proposition 8.13 of [5] and, independently, Gaboriau [22]. Of course, this is to be contrasted with the surface and toral types where the leaves are quasi-isometric to Euclidean space (of dimension 1 and > 1 respectively).

5.4. *Remarks on the proof of the classification theorem*

We now briefly describe the proof of Theorem 5.1. For details, see Section 7 of [5]. As mentioned in the introduction, there is an alternative approach developed by Gaboriau, Levitt, and Paulin.

There is an algorithm for transforming a given band complex (say with a minimal lamination) to another one. When there is a free arc, we collapse from a maximal free subarc as described above. This is Process I. When there are no free subarcs, one performs a *sliding* move, called Process II. These are to be repeated producing a sequence of band complexes. There is a notion of complexity (non-negative half-integer valued). The moves never increase complexity, and whenever Process II is followed by Process I (i.e., whenever free subarcs disappear after a collapse) the complexity strictly decreases. It follows that eventually only Process I or only Process II is performed. In the first case the band complex is of thin type, and in the second one argues that it is of surface or of toral type.

6. Stable actions on $\mathbb{R}$-trees

Here is a sample statement that illustrates how the Rips machine can be applied to understand the structure of a finitely presented group that is acting on an $\mathbb{R}$-tree.

THEOREM 6.1. *Suppose that a torsion-free finitely presented group G is acting non-trivially on an $\mathbb{R}$-tree T by isometries and that all arc stabilizers are trivial. Then one of the following holds.*

- *G splits as a non-trivial free product. In this case one can study the free factors by examining the induced action on T. Either this theorem can be applied to a given factor or this factor is a point stabilizer in T.*
- *G is a free abelian group.*
- *G is the fundamental group of a 2-complex X that contains as a subcomplex a compact connected surface S of negative Euler characteristic and $S \cap \overline{X \setminus S}$ is contained in ∂S. The inclusion induced homomorphism $\pi_1(S) \to \pi_1(X) = G$ is injective and each boundary component of S corresponds to an elliptic isometry in T. There is a filling geodesic measured lamination with measure of full support on S disjoint from ∂S ("filling" means that each complementary component is either simply-connected or it contains a boundary component and its fundamental group is $\mathbb{Z}$).*

PROOF. We may assume that the action is minimal by passing to the minimal subtree, as in Proposition 2.9. Let K be a finite complex with $\pi_1(K) = G$ and choose a resolution $f : \widetilde{K} \to T$ as in Section 4.3. Convert K to a band complex X in the same manner as in Example 4.13. Then apply Theorem 5.1 to replace X by a band complex X' in "normal form".

If some component X'_i is of thin type, then X'_i can be assumed to contain a naked band and hence $G = \pi_1(X')$ splits as a non-trivial free product, so the first possibility holds.

If some component X'_i is of toral type, then either we obtain a free product decomposition of G using one of the points in $Fr(X'_i)$ (so the first possibility holds), or $G = \pi_1(X'_i)$ is free abelian, so the second possibility holds.

If some component X'_i is of surface type, then the third possibility holds.

Finally, if the lamination on X' is simplicial, then G acts on the simplicial tree dual to this lamination. The action can only be "freer" than the original action, so from Bass–Serre theory we conclude that the first possibility holds. □

E. Rips presented a proof of the following theorem at the conference at the Isle of Thorns in the summer of 1991. It answers affirmatively the conjecture of Morgan and Shalen.

THEOREM 6.2 (Rips). *If G is a finitely presented group that acts freely by isometries on an $\mathbb{R}$-tree, then G is the free product of free abelian groups and closed surface groups.*

PROOF. Decompose G into the free product of a free group and freely indecomposable factors, and apply Theorem 6.1 to each freely indecomposable factor. □

As indicated earlier, the methods generalize to stable actions. The following is stated as Theorem 9.5 in [5].

THEOREM 6.3. *Let G be a finitely presented group with a stable action on an $\mathbb{R}$-tree T. Then either*

- *G splits over an extension E-by-cyclic where E fixes an arc of T, or*
- *T is a line and G splits over an extension of the kernel of the action by a free abelian group.*

The structure of the group G obtained in theorems above very much depends on the choice of the resolution. Imagine taking a sequence of finer and finer resolutions of the given stable action, as in the discussion preceding Proposition 4.11. Each band complex in the sequence gives rise to a splitting of G (more precisely, to a graph of groups decomposition of G).

QUESTION. Does the sequence of splittings of G "stabilize"? In other words, is there a structure theorem for G that does not depend on the choice of a resolution, but only on the tree?

QUESTION. Does Theorem 6.3 hold if "finitely presented" is replaced by "finitely generated" in the hypotheses?

Theorem 6.2 holds in the setting of finitely generated groups. Zlil Sela answered the above two questions affirmatively in the case that the action has the additional property that the stabilizers of tripods are trivial. This important case often arises in applications.

The structure of the group acting on an $\mathbb{R}$-tree without the assumption of stability is still very much a mystery.

QUESTION. If a finitely presented group G admits a non-trivial isometric action on an $\mathbb{R}$-tree, does it also admit a non-trivial action on a simplicial tree (i.e., does it admit a non-trivial splitting)?

The answer is affirmative if G is a 3-manifold group by the work of Morgan and Shalen (Proposition 2.1 of [39]).

7. Applications

I will now outline a number of applications of the theory of $\mathbb{R}$-trees. The technique can naturally be used in proofs of finiteness and compactness theorems. Surprisingly, as shown by the work of Zlil Sela, $\mathbb{R}$-trees can also be used to derive various structure theorems in group theory. It is impossible to cover all applications to date, so it seems reasonable to restrict this exposition to outlines of the most typical and the most striking applications. There is no discussion of the work of Rips and Sela on JSJ decompositions of finitely presented groups (see [48]), in part because in the meantime simpler proofs of more general theorems have been found [17,19]. There is no doubt, however, that the intuition coming from the theory of $\mathbb{R}$-trees played a key role in this discovery.

7.1. *Compactifying spaces of geometric structures*

Topologists became interested in $\mathbb{R}$-trees with the work of Morgan and Shalen [37] that shed new light and generalized parts of Thurston's Geometrization Theorem. If M is a closed oriented n-manifold, then having a hyperbolic structure on M is equivalent to having a discrete and faithful representation $\pi_1(M) \to Isom_+ \mathbb{H}^n$ into the orientation-preserving isometry group of the hyperbolic n-space, up to conjugation in $Isom_+ \mathbb{H}^n$. For $n = 2$ and M of genus $g \geqslant 2$ the space

$$Hom_{DF}(\pi_1(M), Isom_+ \mathbb{H}^2)/conj$$

of hyperbolic structures on M is the *Teichmüller space* of M. It is known that this space is homeomorphic to Euclidean space of dimension $6g$–6. The automorphism group $Aut(M)$ of M (homeomorphisms of M modulo isotopy, also known as the *mapping class group* of M) naturally acts on it (with finite isotropy groups), so the Teichmüller space is useful in the study of $Aut(M)$ as it plays the role of the classifying space.

An important ingredient of Thurston's theory of surface automorphisms [61,18] is his construction of an equivariant compactification of the Teichmüller space. An ideal point is represented by a transversely measured geodesic lamination on M (measures that differ by a multiple are equivalent).

From the point of view of $\mathbb{R}$-trees, the construction of this compactification comes from the Compactness Theorem (see Section 3.5). An ideal point is represented by a non-trivial and minimal isometric action of $\pi_1(M)$ on an $\mathbb{R}$-tree, with homothetic actions considered equivalent. Further, from Proposition 3.10 we see that the arc stabilizers are cyclic. Recall that an action of $\pi_1(M)$ on an $\mathbb{R}$-tree is small if it is minimal, does not have global fixed points, and all arc stabilizers are cyclic.

That the two approaches are equivalent follows from the following result of Skora [58].

THEOREM 7.1 (Skora [58]). *If M is a closed hyperbolic surface, then any small action of $\pi_1(M)$ on an $\mathbb{R}$-tree is dual to a unique measured geodesic lamination on M.*

PROOF. The proof using the Rips machine is considerably simpler than the original proof. We focus on the special case when the action is free; the general case is similar. Let $f : \widetilde{X} \to T$ be a resolution of the action. Since $\pi_1(M)$ does not contain $\mathbb{Z} \times \mathbb{Z}$, X cannot have toral components, and since it is not freely decomposable, X cannot have simplicial or thin components, and in fact X must have a single surface component (see Theorem 6.1). Thus X can be taken to be a closed surface equipped with a measured geodesic lamination that fills the surface. To finish the proof, we need to argue that f is an exact resolution. If not, f factors as $f = g\tilde{h}$ through another resolution $g : \widetilde{X}' \to T$, where $\tilde{h} : \widetilde{X} \to \widetilde{X}'$ is equivariant and sends leaves to leaves. By the same argument as above, X' can be taken to be a closed surface with a filling measured geodesic lamination. Thus $h : X \to X'$ is a homotopy equivalence that sends leaves to leaves and locally preserves the transverse measure. In the universal cover (identified with the hyperbolic plane), distinct leaves diverge from each other (in at least one direction), and therefore $\tilde{h}$ cannot send distinct leaves to the same leaf. Since $\tilde{h}$ induces a homeomorphism between the circles at infinity and a lamination is determined by the pairs of endpoints at infinity of its leaves, it follows that h can be taken to be a homeomorphism, showing that f is an exact resolution. □

Theorem 7.1 plays a prominent role in J.-P. Otal's proof [40] of Thurston's Double Limit Theorem, which in turn is a key ingredient in the proof of the Hyperbolization Theorem for 3-manifolds that fiber over the circle.

In dimensions $n > 2$ the celebrated Rigidity Theorem of Mostow states that the space of hyperbolic structures on a closed manifold M^n has at most one point, and the construction using the Compactness Theorem is not particularly exciting in that case. However, it is important in Thurston's proof of the Geometrization Theorem to study the space

$$Hom_{DF}\big(G, Isom_+ \mathbb{H}^n\big)/conj$$

where G is the fundamental group of a compact 3-manifold (with boundary) and $n = 3$. In particular, Thurston needed the fact that this space is compact when the 3-manifold is irreducible, aspherical, acylindrical, and atoroidal. In group-theoretic terms, this means that G is torsion-free and does not split over 1, $\mathbb{Z}$, or $\mathbb{Z}^2$.

THEOREM 7.2 [5]. *Suppose G is finitely presented, not virtually abelian, and does not split over a virtually abelian subgroup. Then the space*

$$Hom_{DF}\big(G, Isom_+ \mathbb{H}^n\big)/conj$$

of homotopy hyperbolic structures on G is compact.

PROOF. If the space is not compact, there is a sequence going to infinity. The Compactness Theorem provides a small action of G on an $\mathbb{R}$-tree. Theorem 6.3 then implies that G splits over a virtually abelian subgroup. (Recall that a discrete group of isometries of $\mathbb{H}^n$ is either virtually abelian or it contains F_2.) □

This theorem generalizes earlier work of Thurston, Morgan and Shalen, and Morgan.

7.2. *Automorphism groups of word-hyperbolic groups*

It is the fundamental observation of Paulin [42] that $\mathbb{R}$-trees arise also in the coarse setting of word-hyperbolic groups in the presence of infinitely many automorphisms of the group. The second part of the proof of the following theorem follows from the Rips machine.

THEOREM 7.3. *Suppose G is a word-hyperbolic group such that $Out(G)$ is infinite. Then G splits over a virtually cyclic subgroup.*

PROOF. Let $f_j : G \to G$ be an infinite sequence of pairwise non-conjugate automorphisms. Each f_j produces an isometric action ρ_j of G on its Cayley graph by sending $g \in G$ to the left translation by $f_j(g)$. The Compactness Theorem provides an action of G on an $\mathbb{R}$-tree T. The arc stabilizers of this action are small by Proposition 3.10, so the claim follows from Theorem 6.3. □

We will now assume that G is a torsion-free word-hyperbolic group. It is an open question whether every word-hyperbolic group has a torsion-free subgroup of finite index (or even whether it is residually finite). It is known that there are only finitely many conjugacy classes of finite order elements [27].

For torsion-free G, an almost-converse of Theorem 7.3 holds [36]. If G splits as a free product $G = A * B$ with A and B non-trivial (infinite!), and if one, say A, is non-abelian, then for a fixed nontrivial $a \in A$ the automorphism $f : G \to G$ that restricts to identity on B and to conjugation by a on A represents an element of infinite order in $Out(G)$. The remaining case is $G = \mathbb{Z} * \mathbb{Z} = F_2$, but $Out(F_2) = GL_2(\mathbb{Z})$ has many elements of infinite order. If G splits over $\mathbb{Z}$, say as $G = A *_C B$, with $A \neq C \neq B$ and $C = \langle c \rangle$ infinite cyclic, then there is a *Dehn twist* automorphism $f : G \to G$ that restricts to identity on B and to conjugation by c on A. This represents an element of infinite order in $Out(G)$ as long as A and B are non-abelian. The case not covered by the "almost-converse" is when one of A or B is infinite cyclic. Finally, if G splits as $G = A*_C$ with $C = \langle c \rangle$ infinite cyclic, then the automorphism (*Dehn twist*) that restricts to the identity on A and sends the "stable letter" t to tc has infinite order in $Out(G)$. For a more detailed discussion that includes cases with torsion see [36].

The above paragraph suggests the following sharpening of Theorem 7.3, at least for torsion-free, freely indecomposable word-hyperbolic groups. Call the subgroup of $Aut(G)$ generated by all inner automorphisms and all Dehn twists (with respect to all possible splittings over infinite cyclic subgroups) the *Internal Automorphism Group*, denoted $Int(G)$. It is a normal subgroup of $Aut(G)$. Note that the celebrated theorem of Dehn [15] (see also [32]) that the mapping class group of a closed orientable surface is generated by Dehn twists can be interpreted as saying $Int(G) = Aut(G)$ where G is the fundamental group of the surface. If the surface is allowed to be non-orientable and to have boundary, then the subgroup of the automorphism group (i.e., the homeomorphism group modulo isotopy) generated by Dehn twists has finite index.

THEOREM 7.4 (Rips and Sela [46]). *If G is a torsion-free, freely indecomposable word-hyperbolic group, then the Internal Automorphism Group has finite index in $Aut(G)$.*

The proof introduces a new idea, the *shortening argument*.

PROOF. Fix a finite generating set $\{\gamma_1, \dots, \gamma_k\}$ for G which is closed under taking inverses and for $f \in Aut(G)$ define

$$d(f) = \max_{1 \leqslant i \leqslant k} \| f(\gamma_i) \|,$$

where $\|\cdot\|$ denotes the word length. In each coset of $Int(G)$ in $Aut(G)$ choose an automorphism f with minimal $d(f)$. Assuming that there are infinitely many cosets, we have an infinite sequence of automorphisms $f_1, f_2, \ldots \in Aut(G)$ that represent distinct cosets of $Int(G)$ and each minimizes the function d in its coset. As in the proof of Theorem 7.3, we view each f_j as giving an action ρ_j of G on its Cayley graph. Note that $1 \in G$ is centrally located for ρ_j (or else composing f_j with an inner automorphism would produce a representative of the same coset with smaller d).

After passing to a subsequence, we obtain a limiting action ρ of G on an $\mathbb{R}$-tree T. We will examine this action and argue that for large j the automorphism f_j can be composed with a Dehn twist in such a way that d is reduced.

Let X be a finite band complex and $\phi: \widetilde{X} \to T$ a resolution of ρ. Choose a basepoint $* \in \widetilde{X}$ that maps to the basepoint in T. We can arrange (see Proposition 4.11) that the distance in $\widetilde{X}$ between $*$ and $\gamma_i(*)$ equals the corresponding distance in T.

Since G is word-hyperbolic and so does not contain $\mathbb{Z} \oplus \mathbb{Z}$, X cannot have any toral components. Similarly, X cannot have any thin components, as we are assuming that G is freely indecomposable. Therefore, all components of X are of simplicial or surface type, and the simplicial pieces have infinite cyclic edge stabilizers.

Let us consider the two extreme cases. First suppose that X is a closed surface with a filling geodesic-like measured lamination. It is a fact of surface theory that there is a homeomorphism $h: X \to X$ fixing the basepoint, which can be taken to be a product of Dehn twists, such that the measure of each $h([\gamma_i])$ is arbitrarily small. This fact can be proved by "unzipping" the band complex (i.e., the "train-track", see [18]) until the bands are arbitrarily thin and taking for h a homeomorphism that sends thick bands to thin bands. Now $f_j\pi_1(h)$ is a "shorter" representative of the coset $f_j\, Int(G)$ for sufficiently large j, a contradiction.

Now suppose that X is simplicial. Let T' be the simplicial tree dual to $\widetilde{X}$. For notational simplicity we assume that T'/G is a single edge, corresponding to an amalgamated product decomposition of G (over $\mathbb{Z}$). The basepoint in $\widetilde{X}$ corresponds to a vertex v in T'. Consider an edge e of T' that has v as an endpoint. The stabilizer of e is infinite cyclic. Say $c \in G$ generates this stabilizer. Let A denote the stabilizer of v and B the stabilizer of the other endpoint of e, so that $G = A *_{\langle c \rangle} B$. Also, without loss of generality we can assume that the length of e is 1. The distance between v and $\gamma_i(v)$ in T' (equivalently, in T) is the minimal $2m_i$ such that γ_i is the product of the form $a_0 b_1 a_2 \cdots$ with the a's in A and the b's in B, and with m_i b's.

Fix a large j and consider the translates of the basepoint $1 \in G$ under the generators γ_i with respect to the representation ρ_j. After rescaling by the constant $d_j = d(f_j)$, the word-metric on G restricted to this finite set is close to the metric induced from T (or T') by restricting to the translates of v by the generators.

Let $b \in B$ be one of the b's occurring in the above representations of the γ_i's. The axis of $f_j(c)$ in G and the first half of the geodesic joining 1 and $f_j(b)$ are within 10δ for a length of about d_j and the translation length of $f_j(c)$ is $\ll d_j$. Replace c by c^{-1} if necessary so that $f_j(c)$ translates from $f_j(b)$ towards 1. Choose the (positive) power m so that $f_j(c)^m$ translates the midpoint of $[1, f_j(b)]$ about halfway towards 1.

We now claim that precomposing f_j by the mth power h of the Dehn twist that fixes A and conjugates B by c has the effect of shortening the representative of the coset $f_j\, Int(G)$.

Indeed, write $\gamma_i = a_0 b_1 a_2 \cdots$ so that $f_j h(\gamma_i) = f_j(a_0) f_j(b_1)^{f_j(c)^m} f_j(a_2) \cdots$. The distance between 1 and $f_j h(\gamma_i)$ can be estimated in the usual way:

$$\begin{aligned} d\big(1, f_j h(\gamma_i)\big) &\leqslant d\big(1, f_j(a_0)\big) + d\big(f_j(a_0), f_j(a_0) f_j(b_1)^{f_j(c)^m}\big) \\ &\quad + d\big(f_j(a_0) f_j(b_1)^{f_j(c)^m}, f_j(a_0) f_j(b_1)^{f_j(c)^m} f_j(a_2)\big) + \cdots \\ &= d\big(1, f_j(a_0)\big) + d\big(1, f_j(b_1)^{f_j(c)^m}\big) + d\big(1, f_j(a_2)\big) + \cdots. \end{aligned}$$

The terms of the form $d(1, f_j(a))$ are small compared to d_j (the ratio goes to 0), and the terms $d(1, f_j(b_k)^{f_j(c)^m})$ are approximately $3d_j/2$. Thus the distance $d(1, f_j h(\gamma_i))$ is estimated above by about $3m_i d_j/2$, and this is much less than $2d(1, f_j(\gamma_i))$ (which is about $2m_i d_j$).

The general case (when T' has perhaps more than one orbit of edges, or when X has both surface and simplicial components) is dealt with in the same way; only notation is more involved. □

A version of the theorem can be proved for torsion-free word-hyperbolic groups that are free products using the classical theory of automorphisms of free products [20,21].

The same method has other applications. Recall that a group G is *co-Hopfian* if every injective endomorphism $G \to G$ is surjective. Non-trivial free products are never co-Hopfian. For our purposes, the group $\mathbb{Z}$ is not freely indecomposable (it splits over the trivial group).

THEOREM 7.5 (Sela [52]). *Every freely indecomposable word-hyperbolic group is co-Hopfian.*

PROOF. Let $Inj(G)$ denote the semi-group of injective endomorphisms of G. The idea is to follow the above argument and show that $Aut(G)$ has finite index in $Inj(G)$. The only difference with the situation $Int(G) \subset Aut(G)$ is that $Inj(G)$ is not a group and $Aut(G)$ is not normal in $Inj(G)$, but those features were never used. Finally, note that if $Aut(G)$ has finite index in $Inj(G)$, then a non-trivial power of every $f \in Inj(G)$ is an automorphism, and so $Inj(G) = Aut(G)$. □

Recall that a group G is *Hopfian* if every surjective endomorphism $G \to G$ is an isomorphism. Z. Sela has announced the following result [53]:

THEOREM 7.6. *Every torsion-free word-hyperbolic group is Hopfian.*

The proof uses more elaborate ideas and will not be outlined here.

THEOREM 7.7 (Gromov [27], Sela [52]). *Let Γ be a finitely presented torsion-free freely indecomposable group and let G be a word-hyperbolic group. Then there are only finitely many conjugacy classes of subgroups of G isomorphic to Γ.*

PROOF. First consider the simple case when Γ does not admit any splittings over $\mathbb{Z}$. Then we argue that there can be only finitely many conjugacy classes of monomorphisms

$f : \Gamma \to G$. For suppose that there are infinitely many. Let $f_1, f_2, \ldots : \Gamma \to G$ be an infinite sequence of pairwise non-conjugate monomorphisms. We thus get a sequence of actions ρ_i of Γ on G: $\rho_i(\gamma)$ acts by left translation by $f_i(\gamma)$. By conjugating each f_i we may assume that $1 \in G$ is centrally located with respect to each ρ_i (and with respect to a fixed finite generating set for Γ). Now pass to a subsequence and obtain an action of Γ on an $\mathbb{R}$-tree. As before, this action induces a splitting of Γ over $\mathbb{Z}$.

If Γ admits a splitting over $\mathbb{Z}$, then we could precompose a given monomorphism $\Gamma \to G$ by automorphisms of Γ and obtain an infinite sequence of non-conjugate monomorphisms $\Gamma \to G$. This phenomenon is precisely what the shortening argument is designed to handle. Given a monomorphism $\Gamma \to G$, conjugate it by an element of G and precompose by an automorphism of Γ so as to make $1 \in G$ centrally located and to make the maximal displacement of 1 smallest possible. Now the claim is that there can be only finitely many such minimizing monomorphisms. The proof of the claim is analogous to the proof of Theorem 7.4. If there are infinitely many such monomorphisms, consider the limiting tree and use it to construct an automorphism $h : \Gamma \to \Gamma$ that can be used to shorten representations ρ_j for large j. □

7.3. *Fixed subgroup of a free group automorphism*

Let F_n be the free group of rank n and $f : F_n \to F_n$ an automorphism. Recall that F_n contains free subgroups of infinite rank. The following theorem was conjectured by Peter Scott.

THEOREM 7.8 (Bestvina–Handel). *The rank of the subgroup $Fix(f)$ of elements of F_n fixed by f is at most n.*

The proof in [6] does not use the theory of $\mathbb{R}$-trees. Sela [51] and Gaboriau, Levitt and Lustig [25] found a simpler argument using $\mathbb{R}$-trees. We now outline their ideas.

First, for $k = 1, 2, \ldots$ let g_k be an automorphism conjugate to f^k such that $1 \in F_n$ is centrally located with respect to the representation ρ_k that to $\gamma \in F_n$ associates left translation $F_n \to F_n$ by $g_k(\gamma)$. This conjugation is necessary in order to apply the Compactness Theorem, but of course $Fix(g_k)$ is in general different from $Fix(f)$. It is therefore more natural to consider elements of F_n fixed up to conjugacy. If f has finite order as an outer automorphism, the rescaling constants remain bounded. Such automorphisms were handled by Culler [12] who showed that the fixed subgroup is either cyclic or a free factor. For non-periodic automorphisms, we analyze the action of F_n on the $\mathbb{R}$-tree T obtained as the limit of a subsequence of representations ρ_k above.

The key observation is that any $\gamma \in F_n$ which is fixed up to conjugacy by f is elliptic in T. Indeed, the translation length of γ can be computed as the limit of ratios

$$\frac{\text{translation length of } g_k(\gamma)}{\text{rescaling factor for } \rho_k}$$

and this converges to 0 since the denominators go to infinity, while the numerators are constant (and equal to the length of the conjugacy class of γ). The same argument shows that periodic conjugacy classes are elliptic in T (and also those that grow slower than the fastest growing conjugacy classes).

Second, we construct a bilipschitz homeomorphism $H : T \to T$ which is equivariant with respect to f, i.e., $h(\gamma(x)) = f(\gamma)(h(x))$. This construction is due to Sela who used it extensively. He calls it the "basic commutative diagram". First form the group $G = F_n \rtimes_f \mathbb{Z} = \langle F_n, t \mid tgt^{-1} = f(g) \rangle$, the mapping torus of f. Each action ρ_k extends to an action $\tilde{\rho}_k$ of G on F_n by sending t to the conjugate of f by the same element used to conjugate f^k. Of course, the extended action is not isometric, only bilipschitz. Pass to a subsequence as usual to obtain a bilipschitz action of G on an $\mathbb{R}$-tree. Restricting to F_n gives the discussion of the first paragraph, while $t \in G$ provides the desired bilipschitz homeomorphism $H : T \to T$.

Third, we promote H to a homothety. This is not absolutely necessary here, but in other applications it comes handy. The following construction is due to Paulin [43]. The Compactness Theorem implies that the space $\mathcal{PED}_0$ of projectivized nontrivial 0-hyperbolic equivariant distance functions on F_n is compact. The preimage of the closed subset $\mathcal{PED}_0^T$ of $\mathcal{PED}_0$ consisting of those projective classes of distance functions d with the property that

$$(x \cdot y)_T \geqslant (x \cdot z)_T \Rightarrow (x \cdot y)_d \geqslant (x \cdot z)_d$$

in $\mathcal{ED}$ is a convex cone: If d_1 and d_2 are two 0-hyperbolic distance functions satisfying the above condition, then $sd_1 + (1 - s)d_2$ is also such a distance function for $0 \leqslant s \leqslant 1$. Subscripts T and d above indicate the metric with respect to which $(\cdot, \cdot)$ is taken. It easily follows that $\mathcal{PED}_0^T$ is a compact absolute retract, and therefore has the fixed point property. By pulling back, H induces a homeomorphism of $\mathcal{PED}_0^T$. A fixed point of H determines a new 0-hyperbolic distance function on F_n with respect to which H is a homothety. By Connecting the Dots (Lemma 2.13), we obtain a new tree that we continue to denote by T. We remark that the new tree may not be homeomorphic to the old, but is rather obtained from the old by collapsing some subtrees. What is important is that arc stabilizers in the new tree are contained in the arc stabilizers of the old tree.

Alternatively, steps 1–3 could have been avoided by quoting some of the theory developed in [6]. See [33], where this alternative construction is carried out in detail.

We now arrive at the heart of the argument.

PROPOSITION 7.9. *Assume that F_n acts on an $\mathbb{R}$-tree T and the action is small. Then all vertex stabilizers of T have rank $\leqslant n$. Further, if there is a vertex stabilizer V of rank n, then the action is simplicial, all edge stabilizers are infinite cyclic, and every vertex stabilizer that is not infinite cyclic is conjugate to V.*

Before giving the proof of Proposition 7.9 we finish the proof of Theorem 7.8. We have seen above that each $\gamma \in Fix(f)$ is elliptic in T. It is an exercise to show that there is a point $v \in T$ fixed by each $\gamma \in Fix(f)$. The ingredients are

(1) the product of two elliptic isometries of T with disjoint fixed point sets is hyperbolic, and
(2) arc stabilizers of T are cyclic.

We may assume that rank($Fix(f)$) > 1, and then v is unique. Since f leaves $Fix(f)$ invariant, equivariance forces H to fix v. In particular, H induces an automorphism $f_v : Stab(v) \to Stab(v)$. If rank($Stab(v)$) $< n$, we can apply induction on the rank and conclude that rank($Fix(f)$) $=$ rank($Fix(f_v)$) $< n$. If rank($Stab(v)$) $= n$ and f_v has finite order (as an outer automorphism) we can apply Culler's result to conclude that rank($Fix(f)$) $\leqslant n$. If rank($Stab(v)$) $= n$ and f_v has infinite order, we can repeat the construction with f_v in place of f. We obtain a sequence of automorphisms $f = f_0, f_v = f_1, f_2, \ldots$. We can stop when the rank of the vertex stabilizer is $< n$ or when the restriction of the automorphism to the vertex stabilizer has finite order. It remains to argue that the sequence must terminate. The tree T constructed above provides a graph of groups decomposition $\mathcal{G}_0$ of F_n with cyclic edge groups (according to Proposition 7.9). The only vertex group is $Stab(v)$ and it has rank n. The next iteration provides a graph of groups decomposition $\mathcal{G}_1$ of $Stab(v)$ of the same nature. We claim that $\mathcal{G}_1$ can be used to refine $\mathcal{G}_0$, by "blowing up" the vertex. Indeed, this will be possible if all edge groups of $\mathcal{G}_0$ are elliptic in $\mathcal{G}_1$. But the edge groups of $\mathcal{G}_0$ are permuted (up to conjugacy) by f, since the orbits of edges are permuted by H, and the claim follows from the observation above that f-periodic conjugacy classes are elliptic in T. If the sequence of automorphisms does not terminate, then continuing in this fashion we obtain graph of groups decompositions of F_n with all edge groups cyclic, with only one vertex, and with more and more edges. This is not possible, for example by the generalized accessibility theorem of [4], or better yet, by abelianizing there can be at most n edges.

PROOF OF PROPOSITION 7.9. Inductively, we assume that Proposition 7.9 holds for free groups of rank $< n$. If $Stab(v)$ is contained in a proper free factor of F_n, then the statement follows inductively on the rank of the underlying free group.

CLAIM 1. *If T is simplicial then all vertex stabilizers have rank $\leqslant n$. If there is a vertex stabilizer of rank n, then all other vertex stabilizers have rank* 1 *and all edge stabilizers are infinite cyclic.*

An edge of T with trivial stabilizer induces a free factorization of F_n which implies the claim by induction. So we can assume that all edge stabilizers are infinite cyclic. Now construct the graph of spaces associated with the graph of groups T/F_n as in [59]. Every vertex in T/F_n is represented by a rose, and every edge by an annulus. Since adding annuli does not change the Euler characteristic, we see that the Euler characteristic of the resulting space, which must be $1-n$, is equal to the sum $\sum(1-r_i)$, where $r_1, r_2, \ldots$ denote the ranks of the vertex labels in T/F_n. Since $r_i > 0$ for each i by assumption, Claim 1 follows.

Assume now rank($Stab(v)$) $< \infty$ and let $\widetilde{X} \to T$ be a resolution of T such that a compact set K in some complementary component $D \subset X$ satisfies $im[\pi_1(K) \to \pi_1(X)] = Stab(v)$ (see Proposition 4.11). Since F_n does not contain $\mathbb{Z} \times \mathbb{Z}$ nor an extension of $\mathbb{Z} \times \mathbb{Z}$ by $\mathbb{Z}$, X cannot have any toral components. Likewise, if X has a thin component, we can transform it so that there is a naked band disjoint from K and we conclude that $Stab(v)$ is contained in a proper free factor. Therefore, X consists of surface and simplicial components. If there is at least one surface component (of negative Euler characteristic), then the Euler characteristic count of Claim 1 implies rank($Stab(v)$) $< n$. So assume that all components of X are simplicial, and let T' be the dual (simplicial) tree. If an edge stabilizer in T' is trivial, then all vertex stabilizers in T', including $Stab(v)$, have rank $< n$. The following claim concludes the proof in case $\text{rank}(Stab(v)) < \infty$:

CLAIM 2. *If all edge stabilizers in T' are infinite cyclic, then T is simplicial.*

To prove the claim, for each primitive $a \in F_n$ consider the subtree $T'_a \subset T'$ consisting of points fixed by a power of a. First note that these subtrees are finite. Indeed, if e is an edge in T'_a whose stabilizer is $\langle a^k \rangle$, then exactly k edges in T'_a (namely, the translates of e by a) can project to the same edge in the quotient. By $T_a \subset T$ denote the image of T'_a under the natural map $\pi' : T' \to T$ (see Section 4.4). Then T_a is a finite tree (by "local injectivity" – see Section 4.4). Moreover, by the equivariance of π, T_a is fixed pointwise by a power of a. It then follows that if a and b are primitive elements with $a \neq b^{\pm 1}$, then T_a and T_b can intersect in at most a point. Since there are only finitely many orbits of the T_a's, the claim follows.

It remains to rule out the possibility that $\text{rank}(Stab(v)) = \infty$. Choose a free factor H of $Stab(v)$ with $n < \text{rank}(H) < \infty$. Let $\widetilde{X} \to T$ be a resolution such that H is in the image of $\pi_1(D) \to \pi_1(X)$ for a complementary component D that corresponds to the orbit of v. Then $im[\pi_1(D) \to \pi_1(X)]$ is contained in $Stab(v)$ and contains H, so that its rank is $> n$. Now analyze the components of X in a similar way as above to reach a contradiction. □

For more details see Gaboriau and Levitt [24]. They also bound the number of orbits of branch points and their "valences" for small actions of F_n.

7.4. *The topology of the boundary of a word-hyperbolic group*

Let G be a word-hyperbolic group and ∂G its boundary. The following theorem was the motivating goal of [7].

THEOREM 7.10. *If G has one end, then ∂G is connected and locally connected.*

The first part of the conclusion (that ∂G is connected) was proved in [7], but the second was proved only under the assumption that ∂G contains no cut points. The theory of $\mathbb{R}$-trees was used to establish:

THEOREM 7.11 (Bowditch, Swarup). *If G has one end, then ∂G contains no cut points.*

SKETCH OF PROOF. For every compact metric space M, Bowditch [9] constructs a canonical map $M \to D$ to a dendrite D. A compact metric space is a *dendrite* if it is locally connected and each pair of points x, y is joined by a unique arc, denoted $[x, y]$. This is done as follows. Two points $x, y \in M$ are NOT equivalent if there is a collection C of cut points in M that each separate x from y and which is order-isomorphic to the rationals. Bowditch argues that the quotient space D is a dendrite. Apply this construction to $M = \partial G$. Since G acts on ∂G, there is an induced action of G on D. If ∂G has a cut point, then it contains a lot of cut points (translates of the original), and Bowditch argues that D is not a point. Further, he shows that the action of G on $T = D \setminus \{endpoints\}$ has trivial arc stabilizers and is non-nesting, in the sense that if J is an arc in T and $g(J) \subseteq J$, then $g(J) = J$ (and hence $g = 1$). The tree T is homeomorphic to an $\mathbb{R}$-tree, but there is no reason why there should be an equivariant $\mathbb{R}$-tree metric on T. If there were, we could apply Theorem 6.3 and conclude that G splits over a 2-ended group. This is where Sacksteder's theorem comes in. We can construct a resolution $\widetilde{X} \to T$ as before, but the lamination on X will not have a transverse measure. Theorem 4.8 provides a transverse measure (perhaps not of full support). It is easy to see that the arc stabilizers of the dual $\mathbb{R}$-tree are trivial. Thus G splits over a 2-ended group.

The proof was completed by Swarup [60]. The idea is to keep refining the splitting as in the proof of Theorem 7.8. So suppose inductively that $\mathcal{G}$ is a graph of groups decomposition of G with 2-ended edge groups. If E is an edge group, the endpoints of the axis of an element of E are identified in the dendrite D [8]. Each vertex group is word-hyperbolic and it is quasi-convex in G. Let $\Lambda(V)$ denote the limit set of a vertex group V of $\mathcal{G}$. It can be argued [8] that for at least one vertex group V the image of $\Lambda(V)$ in D is not a single point. It follows that the induced action of V on T is nontrivial, has trivial arc stabilizers, and all edge groups contained in V are elliptic. Now apply Sacksteder's theorem again to replace T by an $\mathbb{R}$-tree T' on which V acts non-trivially by isometries and with trivial arc stabilizers. The important point is that it can be arranged that the edge groups in V remain elliptic in T'. We then obtain a nontrivial splitting of V over two-ended groups that can be used to refine the graph of groups decomposition $\mathcal{G}$.

The final contradiction comes from the generalized accessibility theorem of [4] that provides an upper bound to the number of edges of a reduced graph of groups decomposition of a finitely presented group with small edge groups. A graph of groups is *reduced* if for each valence one and two vertex of the graph, the label of the vertex properly contains the labels of incident edges. This theorem, combined with the fact that in word-hyperbolic groups the ascending chain condition holds for 2-ended subgroups, implies that from some point on in the construction the refinement of the graph of groups consists of the introduction of new valence two vertices and edges whose labels are finite index subgroups in the old edge labels. Using a technique of Dunwoody [16], one now shows that G splits over a finite subgroup, contradiction. Indeed, the sequence of graphs of groups gives rise to a sequence of pairwise disjoint tracks in a finite complex representing G. The tracks must eventually be parallel to previous tracks. Thus one of the tracks accounts for infinitely many edges in the splitting, i.e., the image of its fundamental group in G must be contained in infinitely many edge groups. But then this image is contained in the intersection, which is finite, and hence the splitting of G over a finite subgroup. □

References

[1] R. Alperin and H. Bass, *Length functions of group actions on λ-trees*, Annals of Math. Stud., Vol. 111, S.M. Gersten and J.R. Stallings, eds, Princeton Univ. Press (1987).

[2] M. Bestvina, *Degenerations of the hyperbolic space*, Duke Math. J. **56** (1988), 143–161.

[3] M. Bestvina and M. Feighn, *Outer limits*, Preprint (1992), revised (1994).

[4] M. Bestvina and M. Feighn, *Bounding the complexity of simplicial group actions on trees*, Invent. Math. **103** (1991), 449–469.
[5] M. Bestvina and M. Feighn, *Stable actions of groups on real trees*, Invent. Math. **121** (1995), 287–321.
[6] M. Bestvina and M. Handel, *Train tracks and automorphisms of free groups*, Ann. Math. **135** (1992), 1–51.
[7] M. Bestvina and G. Mess, *The boundary of negatively curved groups*, J. Amer. Math. Soc. **4** (1991), 469–481.
[8] B.H. Bowditch, *Boundaries of strongly accessible groups*, Preprint (1996).
[9] B.H. Bowditch, *Treelike structures arising from continua and convergence groups*, Mem. Amer. Math. Soc., to appear.
[10] M.R. Bridson and G.A. Swarup, *On Hausdorff–Gromov convergence and a theorem of Paulin*, Enseign. Math. **40** (1994), 267–289.
[11] A. Casson and S. Bleiler, *Automorphisms of surfaces after Nielsen and Thurston*, LMS Student Texts, Vol. 9 (1988).
[12] M. Culler, *Finite groups of outer automorphisms of a free group*, Contemp. Math. **33** (1984), 197–207.
[13] M. Culler and J. Morgan, *Group actions on* $\mathbb{R}$*-trees*, Proc. London. Math. Soc. **55** (1987), 571–604.
[14] W. Dicks and M.J. Dunwoody, *Groups Acting on Graphs*, Cambridge Studies in Advanced Mathematics, Vol. 17, Cambridge University Press (1989).
[15] M. Dehn, *Die Gruppe der Abbildungsklassen*, Acta Math. **69** (1938), 135–206.
[16] M.J. Dunwoody, *The accessibility of finitely presented groups*, Invent. Math. **81** (1985), 449–457.
[17] M.J. Dunwoody and M. Sageev, *JSJ-splittings for finitely presented groups over slender groups*, Preprint.
[18] A. Fathi, F. Laudenbach and V. Poenaru, *Travaux de Thurston sur les surfaces*, Asterisque **66** (1979).
[19] K. Fujiwara and P. Papasoglu, *JSJ decompositions of finitely presented groups and complexes of groups*, Preprint.
[20] D.I. Fouxe-Rabinovich, *Über die Automorphismengruppen der freien Produkte* I, Mat. Sbornik **8** (1940), 265–276.
[21] D.I. Fouxe-Rabinovich, *Über die Automorphismengruppen der freien Produkte* II, Mat. Sbornik **9** (1941), 183–220.
[22] D. Gaboriau, *Dynamique des systèmes d' isométries: Sur les bouts des orbites*, Invent. Math. **126** (1996), 297–318.
[23] E. Ghys and P. de la Harpe (eds.), *Sur les groupes hyperboliques d'après Mikhael Gromov*, Progress in Math., Vol. 88, Birkhäuser (1990).
[24] D. Gaboriau and G. Levitt, *The rank of actions on real trees*, Ann. Sci. École Norm. Sup. **28** (4) (1995), 549–570.
[25] D. Gaboriau, G. Levitt and M. Lustig, *A dendrological proof of the Scott conjecture for automorphisms of free groups*, Proc. Edinburgh Math. Soc. (2) **41** (2) (1998), 325–332.
[26] D. Gaboriau, G. Levitt and F. Paulin, *Pseudogroups of isometries of* $\mathbb{R}$ *and Rips' theorem on free actions on* $\mathbb{R}$*-trees*, Israel J. Math. **87** (1994), 403–428.
[27] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, MSRI Publ., Vol. 8, S. Gersten, ed., Springer, Berlin (1987), 75–263.
[28] G. Levitt, *Constructing free actions on* $\mathbb{R}$*-trees*, Duke Math. J. **69** (1993), 615–633.
[29] G. Levitt, *La dynamique des pseudogroupes de rotations*, Invent. Math. **113** (1993), 633–670.
[30] G. Levitt, *Non-nesting actions on real trees*, Bull. London Math. Soc. **30** (1998), 46–54.
[31] G. Levitt and F. Paulin, *Geometric group actions on trees*, Amer. J. Math. **119** (1997), 83–102.
[32] W.B.R. Lickorish, *A finite set of generators for the homeotopy group of a 2-manifold*, Proc. Camb. Philos. Soc. **60** (1964), 769–778.
[33] M. Lustig, *Automorphisms, train tracks, and non-simplicial* $\mathbb{R}$*-tree actions*, Preprint.
[34] G.S. Makanin, *Equations in a free group*, Math. USSR Izvestiya **21** (1983), 145–163.
[35] R. Martin, *Non-uniquely ergodic foliations of thin type*, Ergodic Theory Dynam. Systems **17** (1997), 667–674.
[36] C.F. Miller, W.D. Neumann and G.A. Swarup, *Some examples of hyperbolic groups*, Preprint.
[37] J. Morgan and P. Shalen, *Valuations, trees, and degenerations of hyperbolic structures*, I, Ann. of Math. (2) **120** (1984), 401–476.
[38] J. Morgan and P. Shalen, *Degenerations of hyperbolic structures*, II: *Measured laminations in 3-manifolds*, Ann. of Math. (2) **127** (1988), 403–456.

[39] J. Morgan and P. Shalen, *Degenerations of hyperbolic structures*, III: *Actions of 3-manifold groups on trees and Thurston's compactness theorem*, Ann. of Math. (2) **127** (1988), 457–519.
[40] J.-P. Otal, *Le théorème d'hyperbolisation pour les variétés fibrées de dimension* 3, Astérisque **235** (1996).
[41] F. Paulin, *Topologie de Gromov équivariant, structures hyperboliques et arbres réels*, Invent. Math. **94** (1988), 53–80.
[42] F. Paulin, *Outer automorphisms of hyperbolic groups and small actions on $\mathbb{R}$-trees*, Arboreal Group Theory, MSRI Publ., Vol. 19, R.C. Alperin, ed., Springer, Berlin (1991), 331–343.
[43] F. Paulin, *Sur les automorphismes extérieurs des groupes hyperboliques*, Ann. Sci. École Norm. Sup. **30** (1997), 147–167.
[44] F. Paulin, *Actions de groupes sur les arbres*, Séminaire Bourbaki 1995/96, Exp. No. 808 (3), Astérisque, no. 241 (1997), 97–137.
[45] A.A. Razborov, *On systems of equations in a free group*, Math. USSR Izvestiya **25** (1985), 115–162.
[46] E. Rips and Z. Sela, *Structure and rigidity in hyperbolic groups I*, Geom. Funct. Anal. **4** (1994), 337–371.
[47] E. Rips and Z. Sela, *Canonical representatives and equations in hyperbolic groups*, Invent. Math. **120** (1995), 489–512.
[48] E. Rips and Z. Sela, *Cyclic splittings of finitely presented groups and the canonical JSJ decomposition*, Ann. Math **146** (1997), 53–109.
[49] R. Sacksteder, *Foliations and pseudogroups*, Amer. J. Math. **87** (1965), 75–102.
[50] Z. Sela, *The isomorphism problem for hyperbolic groups* I, Ann. Math. **141** (1995), 217–283.
[51] Z. Sela, *The Nielsen–Thurston classification and automorphisms of a free group* I, Duke Math. J. **84** (1996), 379–397.
[52] Z. Sela, *Structure and rigidity in (Gromov) hyperbolic groups and discrete groups in rank* 1 *Lie groups*, II, Geom. Funct. Anal. **7** (1997), 561–593.
[53] Z. Sela, *Endomorphisms of hyperbolic groups* I: *The Hopf property*, Preprint.
[54] Z. Sela, *The Nielsen–Thurston classification and automorphisms of the free group* II, Preprint.
[55] J.P. Serre, *Trees*, Springer, Berlin (1980).
[56] P. Shalen, *Dendrology of groups*: *An introduction*, Essays in Group Theory, MSRI Publ., Vol. 8, S.M. Gersten, ed., Springer, Berlin (1987), 265–319.
[57] P. Shalen, *Dendrology and its applications*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific (1991).
[58] R. Skora, *Splittings of surfaces*, J. Amer. Math. Soc. **9** (1996), 605–616.
[59] G.P. Scott and C.T.C. Wall, *Topological methods in group theory*, Homological Group Theory, Lecture Notes, Vol. 36, C.T.C. Wall, ed., London Mathematical Society (1979), 137–203.
[60] G.A. Swarup, *On the cut point conjecture*, Electron. Res. Announc. Amer. Math. Soc. **2** (1996), 98–100.
[61] W. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. **19** (1988), 417–431.

CHAPTER 3

Geometric Structures on 3-Manifolds*

Francis Bonahon

Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113, USA

Contents

*This chapter was written while the author was visiting the Centre Émile Borel, the Institut des Hautes Études Scientifiques and the California Institute of Technology. He would like to thank these institutions for their hospitality. He was also partially supported by NSF grants DMS-9504282 and DMS-9803445.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

In 1977, W.P. Thurston stunned the world of low-dimensional topology by showing that 'many' (in a precise sense) compact 3-dimensional manifolds admitted a unique hyperbolic structure. Of course, hyperbolic 3-manifolds had been around since the days of Poincaré, as a subfield of complex analysis. Work of Andreev [6,7] in the mid-sixties, Riley [112] and Jørgensen [65] in the early seventies, had provided hyperbolic 3-manifolds of increasing complex topology. In a different line of inquiry, W. Jaco and P. Shalen had also observed in the early seventies that the fundamental groups of atoroidal 3-manifolds shared many algebraic properties with those of hyperbolic manifolds. However, the fact that hyperbolic metrics on 3-manifolds were so common was totally unexpected, and their uniqueness had far reaching topological consequences.

At about the same time, the consideration of the deformations and degenerations of hyperbolic structures on non-compact 3-manifolds led Thurston to consider other types of geometric structures. Building on the existing topological technology of characteristic splittings of 3-manifold, he made the bold move of proposing his Geometrization Conjecture which, if we state it in loose terms, says that a 3-manifold can be uniquely decomposed into pieces which each admit a geometric structure. For the so-called Haken 3-manifolds, which at the time were essentially the only 3-manifolds which the topologists were able to handle, this Geometrization Conjecture was a consequence of Thurston's original Hyperbolization Theorem. But for non-Haken 3-manifolds, the conjecture was clearly more ambitious, for instance because it included the Poincaré Conjecture on homotopy 3-spheres as a corollary. Nevertheless, the Hyperbolic Dehn Surgery Theorem and, later, the Orbifold Geometrization Theorem provided a proof of many more cases of the Geometrization Conjecture.

This influx of new ideas completely revolutionized the field of 3-dimensional topology. In addition to the classical arguments of combinatorial topology, many proofs in low-dimensional topology now involve techniques borrowed from differential geometry, complex analysis or dynamical systems. This interaction between topology and hyperbolic geometry has also proved beneficial to the analysis of hyperbolic manifolds and Kleinian groups, where topological insights have contributed to much progress.

Yet, twenty years later, it is still difficult for the non-expert to find a way through the existing and non-existing literature on this topic. For instance, complete expositions of Thurston's Hyperbolization Theorem and of his Orbifold Geometrization Theorem are only beginning to become available. The problem is somewhat different with the topological theory of the characteristic splittings of 3-manifolds. Several complete expositions of the corresponding results have been around for many years, but they are not very accessible because the mathematics involved are indeed difficult and technical.

We have tried to write a reading guide to the field of geometric structures on 3-manifolds. Our approach is to introduce the reader to the main definitions and concepts, to state the principal theorems and discuss their importance and interconnections, and to refer the reader to the existing literature for proofs and details. In particular, there are very few proofs (or even sketches of proof) in this chapter. In a field where unpublished prepublications have historically been very common and important, we tried to only quote references which are widely available, but it was of course difficult to omit such an influential publication as Thurston's original lecture notes [138]. The selection of topics clearly follows the biases of the author, but we also made the deliberate choice of privileging those aspects of geometric structures which have applications to geometric topology. In particular, we eliminated from our discussion the analysis of the geometric properties of infinite volume hyperbolic 3-manifolds, and its relation to complex analysis and complex dynamical systems; we can refer the reader to [18,24,78,79,83,86,138] for some details on this very active domain of research.

1. Geometric structures

1.1. *The case of surfaces*

As an introduction to geometric structures, we first consider a classical property of *surfaces*, namely (differentiable) manifolds of dimension 2. Before going any further, we should mention that we will use the usual implicit convention that a manifold is connected unless specified otherwise; however, submanifolds will be allowed to be disconnected. Also, a manifold will be without boundary, unless it is explicitly identified as a manifold with boundary (or perhaps we should say manifold-with-boundary). Manifolds with boundary will not occur until Section 2.5.

Any (connected) surface S admits a complete Riemannian metric which is locally isometric to the Euclidean plane $\mathbb{E}^2$, the unit sphere $\mathbb{S}^2$ in Euclidean 3-space $\mathbb{E}^3$, or the hyperbolic plane $\mathbb{H}^2$. There are two classical methods to see this: one based on complex analysis, and another one based on the topological classification of surfaces of finite type. We now sketch both, since they each are of independent interest.

Any orientable surface S admits a *complex structure* (or a *Riemann surface structure*), namely an atlas which locally models the surface over open subsets of $\mathbb{C}$, where all changes of charts are holomorphic, and which is maximal for these two properties; see for instance Ahlfors and Sario [2, Chapter III] or Reyssat [110]. The key idea is that, in dimension 2, any Riemannian metric is conformally flat. In other words, if we endow S with an arbitrary Riemannian metric, any point admits a neighborhood which is diffeomorphic to an open subset of $\mathbb{C}$ by an angle preserving diffeomorphism; since the changes of charts respect angles, the Cauchy–Riemann equation then implies that they are holomorphic. Similarly, a possibly non-orientable surface S admits a *twisted complex structure*, defined by a maximal atlas locally modeling S over open subsets of $\mathbb{C}$ and such that all changes of charts are holomorphic or antiholomorphic. This structure lifts to a twisted complex structure on the universal covering $\widetilde{S}$ of S. Since $\widetilde{S}$ is simply connected, we can choose an orientation for it. Then, composing orientation-reversing charts with the complex conjugation $z \mapsto \bar{z}$, we can arrange that all charts are orientation-preserving, so that all changes of charts are holomorphic. We now have a complex structure on $\widetilde{S}$.

The construction of a twisted complex structure on S was only local. The Uniformization Theorem (see [2] or [110] for instance), a global property, asserts that every simply connected complex surface is biholomorphically equivalent to one of the following three surfaces: the complex plane $\mathbb{C}$, the half-plane $\mathbb{H}^2 = \{z \in \mathbb{C};\ \operatorname{Im} z > 0\}$, and the complex projective line $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$. Therefore, $\widetilde{S}$ is biholomorphically equivalent to one of these three surfaces.

The surface S is the quotient of its universal covering $\widetilde{S}$ under the natural action of the fundamental group $\pi_1(S)$. Since the complex structure of $\widetilde{S}$ comes from a twisted complex structure on S, the covering automorphism defined by every element of $\pi_1(S)$ is holomorphic or anti-holomorphic with respect to this complex structure. Also, note that every element of $\pi_1(S)$ acts on $\widetilde{S}$ without fixed points since it is a covering automorphism. We now distinguish cases, according to whether $\widetilde{S}$ is biholomorphically equivalent to $\mathbb{C}$, $\mathbb{H}^2$ or $\mathbb{CP}^1$.

First consider the case where $\widetilde{S}$ is biholomorphically equivalent to $\mathbb{C}$. Every holomorphic or antiholomorphic automorphism of $\mathbb{C}$ is of the form $z \mapsto az + b$ or $z \mapsto c\bar{z} + d$, with a, b, c, $d \in \mathbb{C}$. For a fixed point free automorphism, we must have $a = 1$, or $|c| = 1$ and $c^{\frac{1}{2}}d \in \mathbb{R}$. In particular, every element of $\pi_1(S)$ respects the Euclidean metric of $\widetilde{S}$ coming from the identifications $\widetilde{S} \cong \mathbb{C} \cong \mathbb{E}^2$. This induces on S a metric which, because the metric of $\widetilde{S} \cong \mathbb{E}^2$ is complete, is also complete. Note that this metric on S is *Euclidean*, in the sense that every point of S has a neighborhood which is isometric to an open subset of the Euclidean plane $\mathbb{E}^2$.

Every holomorphic or antiholomorphic automorphism of $\mathbb{H}^2$ is of the form $z \mapsto (az+b)/(cz+d)$ or $z \mapsto (a\bar{z} - b)/(c\bar{z} - d)$, with a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$. Poincaré observed that such an automorphism preserves the *hyperbolic metric* of $\mathbb{H}^2$, defined as the Riemannian metric which at $z \in \mathbb{H}^2$ is $1/\operatorname{Im} z$ times the Euclidean metric of $\mathbb{H}^2 \subset \mathbb{C} \cong \mathbb{E}^2$. (We are here using the topologist's convention for the rescaling of metrics: When we multiply a metric by $\lambda > 0$, we mean that the distances are locally multiplied by λ; in the same situation, a differential geometer would say that the Riemannian metric is multiplied by λ^2.) The metric of $\mathbb{H}^2$ is easily seen to be complete. Therefore, if $\widetilde{S}$ is biholomorphically equivalent to $\mathbb{H}^2$, the hyperbolic metric of $\mathbb{H}^2$ induces a complete metric on $S = \widetilde{S}/\pi_1(S)$. By construction, this metric on S is *hyperbolic*, namely locally isometric to $\mathbb{H}^2$ at each point of S.

Every holomorphic or antiholomorphic automorphism of $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ is of the form $z \mapsto (az+b)/(cz+d)$ or $z \mapsto (a\bar{z}+b)/(c\bar{z}+d)$, with $a, b, c, d \in \mathbb{C}$ and $ad - bc = 1$. In particular, every holomorphic automorphism of $\mathbb{C} \cup \{\infty\}$ has fixed points. It follows that, either the fundamental group $\pi_1(S)$ is trivial, or it is isomorphic to the cyclic group $\mathbb{Z}_2$ and its generator acts antiholomorphically on $\widetilde{S} \cong \mathbb{C} \cup \{\infty\}$. If $z \mapsto (a\bar{z}+b)/(c\bar{z}+d)$ is a fixed point free involution, it is conjugated by a biholomorphic automorphism to the map $z \mapsto -1/\bar{z}$. (Hint: First conjugate it so that it exchanges 0 and ∞.) Therefore, we can choose the holomorphic identification $\widetilde{S} \cong \mathbb{C} \cup \{\infty\}$ so that, either $\pi_1(S)$ is trivial, or $\pi_1(S) \cong \mathbb{Z}_2$ is generated by $z \mapsto -1/\bar{z}$. Identify $\mathbb{C} \cup \{\infty\} = \mathbb{R}^2 \times \{0\} \cup \{\infty\} \subset \mathbb{R}^3 \cup \{\infty\}$ to the unit sphere $\mathbb{S}^2$ of $\mathbb{R}^3$ by stereographic projection. For this identification, the antiholomorphic involution $z \mapsto -1/\bar{z}$ of $\mathbb{C} \cup \{\infty\}$ corresponds to the isometry $x \mapsto -x$ of $\mathbb{S}^2$. Then, the metric of $\widetilde{S} \cong \mathbb{S}^2$ induces a metric on $S = \widetilde{S}/\pi_1(S)$. Note that this metric on S is *spherical*, namely locally isometric to $\mathbb{S}^2$ everywhere, and is necessarily complete by compactness of $\mathbb{S}^2$.

The model spaces $\mathbb{E}^2$, $\mathbb{H}^2$ and $\mathbb{S}^2$ have a property in common: They are all *homogeneous* in the sense that, for any two points in such a space, there is an isometry sending one point to the other. As a consequence, the metrics we constructed on S are *locally homogeneous*: For any two x, $y \in S$, there is a *local isometry* sending x to y, namely an isometry between a neighborhood U of x and a neighborhood V of y which sends x to y. In other words, such a metric locally looks the same everywhere.

We should note that the topology of S is very restricted when $\widetilde{S}$ is biholomorphically isomorphic to $\mathbb{C}$. Indeed, the easy classification of free isometric actions on $\mathbb{E}^2$ shows that S must be a plane, a torus, an open annulus, a Klein bottle, or an open Möbius strip. The topology of S is even more restricted when $\widetilde{S}$ is isomorphic to $\mathbb{CP}^1$: We saw that in this case S must be homeomorphic to the sphere $\mathbb{S}^2$ or to the real projective plane $\mathbb{RP}^2 = \mathbb{S}^2/\{\pm\mathrm{Id}\}$. On the other hand, the case where $\widetilde{S}$ is biholomorphically isomorphic to $\mathbb{H}^2$ covers all the other surfaces. In this 'generic' case, we saw that S admits a complete hyperbolic metric, namely a metric which is locally isometric to the hyperbolic metric of $\mathbb{H}^2$.

Once we know what to look for, there is a more explicit construction of geometric structures, which is based on the topological classification of surfaces of finite type. Recall that a surface has *finite type* if it is diffeomorphic to the interior of a compact surface with (possibly empty) boundary. If S is a surface of finite type, it has a well-defined finite Euler characteristic $\chi(S) \in \mathbb{Z}$.

If $\chi(S) > 0$, the topological classification of surfaces (see for instance Seifert and Threlfall [129, Kapitel 6] or Massey [84, Chapter II]) says that S is a plane, a 2-sphere or a projective plane. In the first case, S is diffeomorphic to the Euclidean plane $\mathbb{E}^2$ and to the hyperbolic plane $\mathbb{H}^2$, and therefore admits a complete Euclidean metric as well as a complete hyperbolic metric. In the remaining two cases, S is diffeomorphic to $\mathbb{S}^2$ or $\mathbb{RP}^2 = \mathbb{S}^2/\{\pm\mathrm{Id}\}$, and therefore admits a (complete) spherical metric.

When $\chi(S) = 0$, S is diffeomorphic to the open annulus, the open Möbius strip, the 2-torus or the Klein bottle. From the classical description of these surfaces as quotients of $\mathbb{E}^2$, we conclude that they all admit a complete Euclidean metric. Considering the quotient of $\mathbb{H}^2 = \{z \in \mathbb{C};\ \mathrm{Im}\, z > 0\}$ by a cyclic group of isometries generated by $z \mapsto \lambda z$ with $\lambda > 1$, or by another cyclic group of isometries generated by $z \mapsto -\lambda\bar{z}$ with again $\lambda > 1$, we can see that the open annulus and the open Möbius strip also admit complete hyperbolic metrics.

Finally, we can consider the case where $\chi(S) < 0$. Then, the classification of surfaces shows that we can find a compact 1-dimensional submanifold γ of S such that each component of $S - \gamma$ is, either a 'pair of pants' (namely an open annulus minus a closed disk) or a 'pair of Möbius pants' (namely an open Möbius strip minus a closed disk).

Consider a closed pair of pants P, namely the complement of three disjoint open disks in the 2-sphere. By an explicit construction involving right angled hexagons in $\mathbb{H}^2$, one can endow P with a hyperbolic metric for which the boundary ∂P is geodesic. In addition, this hyperbolic metric can be constructed so that the length of each boundary component of P can be an arbitrarily chosen positive number (up to isotopy, the hyperbolic metric is actually uniquely determined by the lengths of the boundary components). See [11, Section B.4] for details. In addition, there is a limiting case as we let the length of some boundary components tend to 0, which gives a complete hyperbolic metric on P minus 1, 2 or 3 boundary components, and where the remaining boundary components are still geodesic and of arbitrary lengths; in addition, such a metric has finite area.

There is a similar construction for the closed pair of 'Möbius pants', namely the complement of two disjoint open disks in the projective plane. Such a pair of Möbius pants P can be endowed with a hyperbolic metric for which the boundary is geodesic; in addition the length of each boundary component can be arbitrarily chosen (and there actually is an additional degree of freedom). Again, letting one or two of these lengths tend to 0, one obtains a finite area complete hyperbolic metric on P minus one of two boundary components.

Now, consider an arbitrary surface S of finite type, without boundary and with $\chi(S) < 0$. Using the classification of surfaces, one easily finds a compact 2-sided 1-submanifold C of S such that each component of $S - C$ is, either a pair of pants, or a pair of Möbius pants. For each component S_i of $S - C$, let $\widehat{S}_i$ be the surface with boundary formally obtained by adding to each end of S_i the component of C that is adjacent to it, with the obvious topology. We saw that we can endow each $\widehat{S}_i$ with a complete hyperbolic metric with geodesic boundary. Now, the surface S is obtained from the disjoint union of the $\widehat{S}_i$ by gluing back together the boundary components which correspond to the same component of C. We can choose the hyperbolic metric on the $\widehat{S}_i$ so that, when two boundary components are to be glued back together, they have the same length and the gluing map is an isometry. Then, one easily checks that the resulting metric on S is hyperbolic, even along C.

In this way, we can explicitly endow any surface S of finite type such that $\chi(S) < 0$ with a complete hyperbolic metric with finite area. Note that the isotopy class of this metric is in general far from being unique. Indeed, we were able to freely choose the length of the components of the 1-submanifold C. In a metric of negative curvature, every homotopy class of simple closed curves contains at most one closed geodesic. It follows that if, in the construction, we start from two hyperbolic metrics on the $\widehat{S}_i$ which give different lengths to some boundary components, the resulting two hyperbolic metrics on S cannot be isotopic. There is an additional degree of freedom associated to each component of C: when we glue back together the corresponding boundary components of the $\widehat{S}_i$, we can vary the gluing map by pre-composing it with an orientation-preserving isometry of one of these boundary components. If we add to this the degree of freedom hidden in the Möbius pant components of $S - C$ which we mentioned earlier, this clearly indicates that the hyperbolic metric of S is far from being unique.

However, the *Teichmüller space* $\mathcal{T}(S)$ of S, defined as the space of isotopy classes of all finite area complete hyperbolic metric on S, can be completely analyzed along these lines. In particular, it is homeomorphic to a Euclidean space of dimension $3|\chi(S)| - e$, where e is the number of ends of S. Good references include Benedetti and Petronio [11, Section B.4] or Fathi, Laudenbach and Poenaru [37, Exposé 7].

1.2. *General definitions*

The above analysis of surfaces suggests the following definition. A *geometric structure* on a connected manifold M without boundary is a locally homogeneous Riemannian metric m on M. As usual, the Riemannian metric turns M into a metric space, where the distance from x to y is defined as the infimum of the lengths of all differentiable arcs going from x to y. A geometric structure is *complete* when the corresponding metric space is complete. We just saw that every connected surface admits such a complete geometric structure.

Given a geometric structure m on M, we can always rescale the metric by a constant to obtain a new geometric structure. More generally, we can change m in the following way. For $x \in M$, consider all local isometries φ sending x to itself; the corresponding differentials $T_x\varphi : T_xM \to T_xM$ form a group G_x of linear automorphisms of the tangent space $T_x M$, called the *isotropy group* of the geometric structure at $x \in M$. Note that the isotropy group respects the positive definite quadratic form m_x defined by m on $T_x M$, and is therefore compact. Also, if φ is a local isometry sending x to y, the differential of φ sends the isotropy group of x to the isotropy group of y. The isotropy group of x is therefore independent of x up to isomorphism. If we fix a point $x_0 \in M$, let m'_{x_0} be another positive definite quadratic form on $T_{x_0} M$ which is respected by the isotropy group G_{x_0}. We can then transport m'_{x_0} to any other tangent space $T_x M$ by using the differential $T_x \varphi : T_{x_0} M \to T_x M$ of any local m-isometry φ sending x_0 to x; the fact that G_{x_0} preserves m'_{x_0} guarantees that this does not depend on the choice of φ. We define in this way a new Riemannian metric on M, which is locally homogeneous by construction.

If the isotropy group G_x acts transitively on $T_x M$, the above construction simply yields a rescaling of the metric. For geometric structures with non-transitive isotropy groups, the modifications of the geometric structures can be a little more complex. However, they still do not differ substantially from the original geometric structure. This leads us to consider a weaker form of geometric structures, in order to neutralize these trivial deformations.

A complete geometric structure on M lifts to a complete geometric structure on the universal covering $\widetilde{M}$ of M. A result of Singer [131] asserts that a complete locally homogeneous Riemannian metric on a simply connected manifold is actually homogeneous. In particular, the isometry group of $\widetilde{M}$ acts transitively in the sense that, for every $\widetilde{x}, \widetilde{y} \in \widetilde{M}$, there exists an isometry g of $\widetilde{M}$ such that $g(\widetilde{x}) = \widetilde{y}$. We consequently have a Riemannian manifold $X = \widetilde{M}$ and a group G of isometries of X acting transitively on X. In addition, M admits an atlas $\{\varphi_i : U_i \to V_i\}_{i \in I}$ which locally models M over X and where all changes of charts are restrictions of elements of G. Namely, each φ_i is a diffeomorphism between an open subset U_i of M and an open subset V_i of X, the union of the U_i is equal to M, and each change of charts $\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ is the restriction of an element of G. Finally, note that an isometry g of X is completely determined by the image gx of a point x and by the differential $T_x g : T_x X \to T_{gx} X$. (Hint: Follow the geodesics.) If we endow G with the compact open topology, it follows that for every $x \in X$ the stabilizer $G_x = \{g \in G;\ gx = x\}$ is compact, since it is homeomorphic to its image in the orthogonal group of isometries of the tangent space $T_x M$.

More generally, consider a group G acting effectively[1] and transitively on a connected manifold X, in such a way that the stabilizer G_x of each point $x \in X$ is compact for the compact open topology. An *(X, G)-structure* on a manifold M is defined by an atlas $\{\varphi_i : U_i \to V_i\}_{i \in I}$ which locally models M over X and where all changes of charts are restrictions of elements of G, as defined above. More precisely, such an *(X, G)-atlas* is contained in a unique maximal (X, G)-atlas, and an (X, G)-structure on M is defined as a maximal (X, G)-atlas.

In this situation, the hypothesis that the stabilizers G_x are compact guarantees the existence of a Riemannian metric on X which is invariant under the action of G. Indeed, if we fix a base point $x_0 \in X$, we can average an arbitrary positive definite quadratic form on $T_{x_0} X$ with respect to the Haar measure of G_{x_0} to obtain a G_{x_0}-invariant positive definite quadratic form. If we use $g \in G$ to transport this quadratic form to $T_{gx_0} X$, we now have a well defined Riemannian metric on X which is invariant under the action of G. This metric

[1] Recall that a group G acts *effectively* on a set X if no non-trivial element of G acts by the identity on X.

is homogeneous by construction, and therefore complete. Also, note that this construction establishes a one-to-one correspondence between G-invariant Riemannian metrics on X and G_{x_0}-invariant positive definite quadratic forms on the tangent space $T_{x_0}X$.

If M is endowed with an (X, G)-structure and if we choose a G-invariant Riemannian metric on X, we can pull back the metric of X by the charts of the (X, G)-atlas. This gives a locally homogeneous metric on M, namely a geometric structure on M.

An (X, G)-structure is *complete* if, for an arbitrary choice of a G-invariant metric on X, the associated geometric structure on M is complete. Note that different choices of a G-invariant metric on X give geometric structures on M which are Lipschitz equivalent, so that this notion of completeness is independent of the choice of the G-invariant metric on X.

As a summary, a complete geometric structure on M defines a complete (X, G)-structure on M, where X is the universal covering of M and where G is the isometry group of X. Conversely, a complete (X, G)-structure on M defines a complete geometric structure on M, modulo the choice of a G_{x_0}-invariant positive definite quadratic form on the tangent space on $T_{x_0}X$. So, intuitively, a complete (X, G)-structure corresponds to a metric independent version of a complete geometric structure. The reader should however beware of a few phenomena such as the fact that, if we start from a (X, G)-structure and a G-invariant metric on M, associate to them a complete geometric structure, and then consider the corresponding (X', G')-structure, the final geometric model may be much more symmetric than the original one in the sense that the stabilizers $G'_{x'}$ may be larger than the stabilizers G_x.

A *geometry* consists of a pair (X, G) as above, namely where X is a connected manifold, where the group G acts effectively and transitively on X, and where all stabilizers G_x are compact. This is also equivalent to the data of a connected Lie group G and of a compact Lie subgroup H of G, if we associate to this data the homogeneous space $X = G/H$ endowed with the natural left action of G.

We identify two geometries (X, G) and (X', G') if there is a diffeomorphism from X to X' which sends the action of G to the action of G'. An (X, G)-structure on M naturally lifts to an $(\widetilde{X}, \widetilde{G})$-structure where $\widetilde{G}$ consists of all lifts of elements of G to the universal covering $\widetilde{X}$ of X. Therefore, we can restrict attention to geometries (X, G) where X is simply connected. Also, if the geometry (X, G) can be enlarged to a more symmetric geometry (X, G') with $G \subset G'$, every (X, G)-structure naturally defines an (X, G')-structure. Consequently, if we want to classify all possible geometries in a given dimension, it makes sense to restrict attention to geometries (X, G) which are *maximal*, namely where X is simply connected and where there is no larger geometry (X', G') with $G \subset G'$ and $G \neq G'$.

2. The eight 3-dimensional geometries

We now focus on the dimension 3, and want to list all maximal geometries (X, G) where X is 3-dimensional. As indicated above, this amounts to listing all pairs (G, H) where G is a Lie group, H is a compact Lie subgroup of G, and the quotient G/H has dimension 3 and is simply connected (we let the reader translate the maximality condition into this context). Note that H must be isomorphic to a closed subgroup of $O(3)$. Listing all such geometries now becomes a relatively easy exercise using the Lie group machinery.

However, it is convenient to decrease the list even further. We will see that complete geometric structures of finite volume tend to have better uniqueness properties. Therefore, it makes sense to restrict attention to geometries for which there is at least one manifold admitting a complete (X, G)-structure of finite volume; note that this finite volume property does not depend on the choice of a G-invariant metric on X.

In this context, Thurston observed that there are exactly 8 maximal geometries (X, G) for which there is at least one finite volume complete (X, G)-structure. This section is devoted to a description of these eight geometries and of their first properties. The article by Scott [126] constitutes a very complete reference for this material.

2.1. *The three isotropic geometries*

The three 2-dimensional geometries (X, G) which we encountered are *isotropic* in the sense that, for any two points $x, x' \in X$ and any half-lines $\mathbb{R}^+ v \subset T_x X$ and $\mathbb{R}^+ v' \subset T_{x'} X$ in the tangent spaces of X at x and x', there is an element of G sending x to x' and v to v'. This is equivalent to the property that the stabilizer G_x acts transitively on the set of half-lines in the tangent space $T_x X$. In other words the geometry (X, G) is isotropic if, not only does X look the same at every point, but it also looks the same in every direction.

In dimension 3 (and actually in any dimension), there similarly are three isotropic maximal geometries. If, for an isotropic geometry (X, G), we endow X with a G-invariant Riemannian metric, passing to the orthogonal shows that we can send any plane tangent to X at $x \in X$ to any other plane tangent to X at $x' \in X$ by an element of G. As a consequence, any G-invariant metric on X must have constant sectional curvature. A classical result in differential geometry says that, for every $K \in \mathbb{R}$ and every dimension n, there is only one simply connected complete Riemannian manifold of dimension n and of constant sectional curvature K, up to isometry; see for instance Wolf [154]. Since rescaling the metric by $\lambda > 0$ multiplies the curvature by λ^{-2}, this leaves us with only 3 possible models for X, according to whether the curvature is positive, 0 or negative.

When the curvature is positive, we can rescale the metric so that the curvature is $+1$. Then, X is isometric to the unit sphere

$$\mathbb{S}^3 = \left\{ (x_0, x_1, x_2, x_3) \in \mathbb{R}^4;\ \sum_{i=0}^{3} x_i^2 = 1 \right\}$$

with the Riemannian metric induced by the Euclidean metric of $\mathbb{R}^4 = \mathbb{E}^4$. By maximality, G is equal to the isometry group $\mathrm{Isom}(\mathbb{S}^3)$ of $\mathbb{S}^3$. This isometry group clearly contains the orthogonal group $\mathrm{O}(4)$. Since $\mathrm{O}(4)$ acts transitively on the space of orthonormal frames[2] of $\mathbb{S}^3$, this inclusion is actually an equality, namely $G = \mathrm{Isom}(\mathbb{H}^3) = \mathrm{O}(4)$.

When the curvature is 0, X is isometric to the Euclidean space $\mathbb{E}^3$, with the usual Euclidean metric. Again, G coincides with the isometry group $\mathrm{Isom}(\mathbb{E}^3)$, which is described by the exact sequence

$$0 \to \mathbb{R}^3 \to \mathrm{Isom}(\mathbb{E}^3) \to \mathrm{O}(3) \to 0$$

[2] Recall that an orthogonal frame is an orthonormal basis in the tangent space $T_x\mathbb{S}^3$ of some $x \in \mathbb{S}^3$.

where the subgroup $\mathbb{R}^3$ consists of all translations, and where the map $\mathrm{Isom}(\mathbb{E}^3) \to \mathrm{O}(3)$ is defined by considering the tangent part of an isometry. Any choice of a base point $x_0 \in \mathbb{E}^3$ defines a splitting of this exact sequence, by sending $g \in \mathrm{O}(3)$ to the isometry of $\mathbb{E}^3$ that fixes x_0 and is tangent to g. In particular, this describes $\mathrm{Isom}(\mathbb{E}^3)$ as the semi-direct product of $\mathbb{R}^3$ and of $\mathrm{O}(3)$, twisted by the usual action of $\mathrm{O}(3)$ on $\mathbb{R}^3$.

When the curvature is negative, we can again rescale the metric so that the curvature is -1. Then, X is isometric to the hyperbolic 3-space

$$\mathbb{H}^3 = \{(u, v, w) \in \mathbb{R}^3;\ w > 0\}$$

endowed with the Riemannian metric which, at (u, v, w), is $1/w$ times the Euclidean metric. Among the three isotropic geometries, the geometry of $\mathbb{H}^3$ is probably the least familiar, but it is also the richest. For instance we will see that, as in the case of surfaces, there are many more 3-manifolds which admit a geometry modelled over $\mathbb{H}^3$ than over $\mathbb{E}^3$ or $\mathbb{S}^3$. An isometry of $\mathbb{H}^3$ continuously extends to its closure in $\mathbb{R}^3 \cup \{\infty\}$. The boundary of $\mathbb{H}^3$ in $\mathbb{R}^3 \cup \{\infty\}$ is $\mathbb{R}^2 \times \{0\} \cup \{\infty\}$, which the standard isomorphism $\mathbb{R}^2 \cong \mathbb{C}$ identifies to the complex projective line $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$. It can be shown that any homeomorphism of $\mathbb{C} \cup \{\infty\}$ that is induced by an isometry of $\mathbb{H}^3$ is holomorphic or antiholomorphic, and therefore is of the form $z \mapsto (az + b)/(cz + d)$ or $z \mapsto (a\bar{z} + b)/(c\bar{z} + d)$ with a, b, c, $d \in \mathbb{C}$ with $ad - bc = 1$. Conversely, every holomorphic or antiholomorphic homeomorphism φ of $\mathbb{C} \cup \{\infty\}$ extends to an isometry of $\mathbb{H}^3$. The easier way to see this is probably to remember that such a φ can be written as a product of inversions across circles, to extend an inversion of $\mathbb{C} \cup \{\infty\}$ across the circle C to the inversion of $\mathbb{R}^3 \cup \{\infty\}$ across the sphere that has the same center and the same radius as C, and to check that the inversion across such a sphere respects $\mathbb{H}^3$ and the metric of $\mathbb{H}^3$.

2.2. *The four Seifert type geometries*

In contrast to the dimension 2, there is enough room in dimension 3 to allow maximal geometries (X, G) which are not isotropic. Namely, for such a geometry, there is at each point x a preferred line L_x in the tangent space $T_x X$ such that, for each $g \in G$ and each $x \in X$, the differential $T_x g : T_x X \to T_x X$ sends the line L_x to L_{gx}.

The first two such geometries are provided by the Riemannian manifolds $\mathbb{S}^2 \times \mathbb{E}^1$ and $\mathbb{H}^2 \times \mathbb{E}^1$, endowed with the product metric.

For $X = \mathbb{H}^2 \times \mathbb{E}^1$, say, consider the natural action of the group $G = \mathrm{Isom}(\mathbb{H}^2) \times \mathrm{Isom}(\mathbb{E}^1)$, where $\mathrm{Isom}(Y)$ denotes the isometry group of Y. This action respects the metric of X, and is clearly transitive. Note that, for every $(x, y) \in \mathbb{H}^2 \times \mathbb{E}^1$, the differential of any element of the stabilizer $G_{(x,y)}$ respects the line $L_{(x,y)} = 0 \times T_y\mathbb{E}^1 \subset T_{(x,y)}\mathbb{H}^2 \times \mathbb{E}^1$. Therefore, the geometry (X, G) is non-isotropic, and of the type mentioned above.

It remains to check that this geometry (X, G) is maximal. This is clearly equivalent to the property that, for every G-invariant metric m on X, the isometry group of m cannot be larger than G. At each point $(x, y) \in X$, the metric m must be invariant under the action of the stabilizer $G_{(x,y)}$. In particular, since $G_{(x,y)}$ contains maps which rotate X around $\{x\} \times \mathbb{E}^1$, the bilinear form induced by m on $T_{(x,y)}X$ must be invariant under rotation around $L_{(x,y)} = 0 \times T_y\mathbb{E}^1$. It follows that m must be obtained from the product metric m_0 by rescaling it by a factor of $\lambda_1 > 0$ in the direction of the line $L_{(x,y)} = 0 \times T_y\mathbb{E}^1$ and by a factor of $\lambda_2 > 0$ in the direction of the orthogonal plane $L^{\perp}_{(x,y)} = T_x\mathbb{H}^2 \times 0$ (keeping these two subspaces orthogonal). To show that G is the whole isometry group of such a metric m, note that the sectional curvature of m along a plane $P \subset T_{(x,y)}X$ is 0 if P contains the line $L_{(x,y)}$, is $-\lambda_2^{-2}$ if P is equal to the orthogonal $L^{\perp}_{(x,y)}$, and is strictly between 0 and $-\lambda_2^{-2}$ otherwise. It follows that the differential of every m-isometry φ must send $L_{(x,y)}$ to $L_{\varphi(x,y)}$ and $L^{\perp}_{(x,y)}$ to $L^{\perp}_{\varphi(x,y)}$. In particular, at an arbitrary point $(x_0, y_0) \in X$, there is an isometry $\varphi' \in G$ such that $\varphi'(x_0, y_0) = \varphi(x_0, y_0)$ and $T_{(x_0,y_0)}\varphi' = T_{(x_0,y_0)}\varphi$, which implies that $\varphi = \varphi' \in G$. Therefore, every m-isometry φ is an element of G.

Replacing $\mathbb{H}^2$ by $\mathbb{S}^2$, we similarly prove that the manifold $X = \mathbb{S}^2 \times \mathbb{E}^1$ endowed with the natural action of $G = \mathrm{Isom}(\mathbb{S}^2) \times \mathrm{Isom}(\mathbb{E}^1)$ defines a maximal geometry (X, G) (the only difference being that the sectional curvature along a plane is now between 0 and $+\lambda_2^{-2}$).

Note that the geometry where $X = \mathbb{E}^2 \times \mathbb{E}^1$ and $G = \mathrm{Isom}(\mathbb{E}^2) \times \mathrm{Isom}(\mathbb{E}^1)$ is conspicuously absent. This is because $\mathbb{E}^2 \times \mathbb{E}^1$ is identical to the Euclidean 3-space $\mathbb{E}^3$, and G can therefore be extended to the larger group $\mathrm{Isom}(\mathbb{E}^3)$. Therefore, this geometry is not maximal.

There are also twisted versions of these product geometries. We first describe an explicit model for the twisted product $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$. Let $T^1\mathbb{H}^2$ be the unit tangent bundle of $\mathbb{H}^2$, consisting of all tangent vectors of length 1 of $\mathbb{H}^2$. Consider the natural projection $p: T^1\mathbb{H}^2 \to \mathbb{H}^2$, associating its base point to each $v \in T^1\mathbb{H}^2$.

The metric of $\mathbb{H}^2$ determines a metric on $T^1\mathbb{H}^2$ as follows: The tangent space of $T^1\mathbb{H}^2$ at $v \in T^1\mathbb{H}^2$ naturally splits as the direct sum of a line L_v and of a plane P_v, where L_v is the tangent line to the fiber $p^{-1}(p(v))$, and where P_v consists of all infinitesimal parallel translations of v along geodesics passing through the point $p(v) \in \mathbb{H}^2$. The norm defined by the metric of $\mathbb{H}^2$ on $T_{p(v)}\mathbb{H}^2$ induces a metric on the fiber $p^{-1}(p(v)) \subset T_{p(v)}\mathbb{H}^2$, making it isometric to the unit circle $\mathbb{S}^1$, and this metric induces a norm on the line L_v tangent to $p^{-1}(p(v))$. Also, the restriction of the differential dp_v identifies the plane P_v to the tangent space $T_{p(v)}\mathbb{H}^2$, and the metric of $\mathbb{H}^2$ then defines a norm on P_v. The Riemannian metric of $T^1\mathbb{H}^2$ is defined by the property that, at each $v \in T^1\mathbb{H}^2$, it restricts to the above norms on L_v and P_v and it makes these two spaces orthogonal.

The construction of this metric is intrinsic enough that it is respected by the natural lift $v \mapsto T_{p(v)}\varphi(v)$ of each isometry $\varphi: \mathbb{H}^2 \to \mathbb{H}^2$. It is also respected by the other natural transformations of $T^1\mathbb{H}^2$ that rotate each vector v by a fixed angle θ, for every θ. In particular, this metric makes $T^1\mathbb{H}^2$ a homogeneous Riemannian manifold.

The space $T^1\mathbb{H}^2$ has the homotopy type of a circle. The model for $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ is its universal covering $\widetilde{T}^1\mathbb{H}^2$.

Topologically, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ is homeomorphic to $\mathbb{H}^2 \times \mathbb{E}^1$, by a homeomorphism which conjugates the submersion $\widetilde{p}: \mathbb{H}^2 \widetilde{\times} \mathbb{E}^1 \to \mathbb{H}^2$ lifting p to the projection $\mathbb{H}^2 \times \mathbb{E}^1 \to \mathbb{H}^2$. However, the situation is metrically very different. Indeed, if α is an oriented differentiable curve going from x to itself in $\mathbb{H}^2$ and if v is in the fiber $\widetilde{p}^{-1}(x)$, there is a unique way of lifting α to a curve $\widetilde{\alpha}$ in $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ that begins at v and is everywhere orthogonal to the fibers $p^{-1}(\alpha(t))$. It immediately follows from the Gauss–Bonnet formula that, in contrast to what happens in the case of $\mathbb{H}^2 \times \mathbb{E}^1 \to \mathbb{H}^2$ (where $\widetilde{\alpha}$ returns to its starting point v), the end point of $\widetilde{\alpha}$ sits at a signed distance of $-A$ from v in the fiber $\widetilde{p}^{-1}(x)$, where A is the signed area enclosed by α in $\mathbb{H}^2$ and where $\widetilde{p}^{-1}(x)$ is oriented by the orientation of $\mathbb{H}^2$.

Since the Riemannian manifold $T^1\mathbb{H}^2$ is homogeneous, so is its universal covering $\widetilde{T}^1\mathbb{H}^2 = \mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$. In particular, the isometry group of $X = \mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ contains the group G generated by the vertical translations along the fibers and by the lifts of the isometries of $T^1\mathbb{H}^2$ associated to the isometries of $\mathbb{H}^2$.

It remains to see that the geometry (X, G) so defined is maximal. As usual, it suffices to prove that the isometry group of any G-invariant metric is equal to G. The action of the stabilizer G_v on the tangent space T_vX contains all rotations around the line L_v tangent to the fiber $p^{-1}(p(v))$. Therefore, any G-invariant metric m on X must be obtained by rescaling the original metric by a uniform factor along L_v and by another uniform factor along the plane P_v orthogonal to L_v. A straightforward computation shows that the sectional curvature of such a metric m along a plane $P \subset T_vX$ is maximal when P contains L_v, and is minimal when P is orthogonal to L_v. As a consequence, the differential of every isometry φ of m must send L_v to $L_{\varphi(v)}$, and therefore commutes with the projection $\widetilde{p}: \mathbb{H}^2 \widetilde{\times} \mathbb{E}^1 \to \mathbb{H}^2$. Also, considering the lift of a closed curve α enclosing a non-zero area in $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$, we see that an m-isometry φ respects the orientation of the fibers of $\widetilde{p}$ if and only if the induced map $\mathbb{H}^2 \to \mathbb{H}^2$ is orientation-preserving. At this point, for every m-isometry φ and for an arbitrary $v \in X$, one easily finds an element $\psi \in G$ with $\varphi(v) = \psi(v)$ and $T_v\varphi = T_v\psi$, from which we conclude that $\varphi = \psi$. Therefore, the geometry of $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ with the transformation group G is maximal.

To conclude this discussion of $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$, we should note that the action of the orientation-preserving isometry group of $\mathbb{H}^2$ on $T^1\mathbb{H}^2$ is transitive and free, so that the choice of a base point identifies this group to $T^1\mathbb{H}^2$. We saw in Section 1.1 that every orientation-preserving isometry of $\mathbb{H}^2$ is a linear fractional map of the form $z \mapsto (az + b)/(cz + d)$ with a, b, c, $d \in \mathbb{R}$, which we can normalize so that $ad - bc = 1$. Associating to such a linear fractional the matrix with coefficients a, b, c, d defines a group isomorphism between the orientation-preserving isometry group of $\mathbb{H}^2$ and the matrix group $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm \mathrm{Id}\}$. For this reason, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1 = \widetilde{T}^1\mathbb{H}^2$ is often denoted by $\widetilde{\mathrm{PSL}}_2(\mathbb{R})$.

There is a similar twisted product $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$. An explicit model for $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ can be constructed as follows. Let $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ be $\mathbb{R}^3$ with the Riemannian metric

$$ds^2 = dx^2 + dy^2 + \left(dz - \tfrac{1}{2}y\,dx + \tfrac{1}{2}x\,dy\right)^2.$$

There is a Riemannian submersion $p: \mathbb{E}^2 \widetilde{\times} \mathbb{E}^1 \to \mathbb{E}^2$ defined by $p(x, y, z) = (x, y)$. Any isometry $\varphi: \mathbb{E}^2 \to \mathbb{E}^2$ lifts to an isometry Φ of $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ defined by the formula

$$\Phi(x, y, z) = \left(\varphi(x, y), \varepsilon z + \tfrac{1}{2}bx - \tfrac{1}{2}ay\right),$$

where $\varepsilon = +1$ or -1 according to whether φ preserves or reverses the orientation of $\mathbb{E}^2$ and where $(a, b) = \varphi(0, 0)$. Also, every vertical translation of $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1 = \mathbb{R}^3$ is an isometry. It follows that the Riemannian manifold $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ is homogeneous.

The Riemannian submersion $p: \mathbb{E}^2 \widetilde{\times} \mathbb{E}^1 \to \mathbb{E}^2$ is 'twisted' in a way which is very similar to the one we observed for $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$: If α is a curve going from x to x in $\mathbb{E}^2$ and if we lift α to a curve $\widetilde{\alpha}$ in $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ that is everywhere orthogonal to the fibers, the end point of $\widetilde{\alpha}$

is at oriented distance $-A$ from its starting point in the fiber $p^{-1}(x)$, where A is the signed area enclosed by α in $\mathbb{E}^2$; this immediately follows from the expression of the area A as the line integral of $\frac{1}{2}x\,dy - \frac{1}{2}y\,dx$ along α.

The manifold $X = \mathbb{E}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$, endowed with the action of the group G generated by all vertical translations and by the lifts of isometries of $\mathbb{E}^2$ described above, defines a geometry (X, G). The fact that this geometry is maximal is proved by the same methods as for $\mathbb{H}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$.

The isometry group G of $X = \mathbb{E}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$ is easily seen to be nilpotent. It is the only nilpotent group among the groups associated to maximal 3-dimensional geometries. For this reason, the geometry of $X = \mathbb{E}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$ is often called the Nil geometry.

We could also expect a similarly twisted geometry $\mathbb{S}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$. However, a computation shows that the sectional curvature of such a homogeneous manifold would have to be everywhere positive, and the model space would consequently have to be compact. There is a twisted product $\mathbb{S}^2 \mathbin{\tilde{\times}} \mathbb{S}^1$ corresponding to the universal cover $\widetilde{T}^1\mathbb{S}^2$ of the unit tangent bundle of $\mathbb{S}^2$, as in the case of $\widetilde{T}^1\mathbb{H}^2 = \mathbb{H}^2 \mathbin{\tilde{\times}} \mathbb{E}^1$. Note that each fiber of $\widetilde{T}^1\mathbb{S}^2 = \mathbb{S}^2 \mathbin{\tilde{\times}} \mathbb{S}^1$ double covers a fiber of $T^1\mathbb{S}^2$. In this geometry, if we lift a closed curve α in $\mathbb{S}^2$ to a curve $\tilde{\alpha}$ in $\mathbb{S}^2 \mathbin{\tilde{\times}} \mathbb{S}^1$ which is everywhere orthogonal to the $\mathbb{S}^1$ factor, the end points of $\tilde{\alpha}$ are $+A$ apart in the fiber above the starting point of α, where A is the area enclosed by α in $\mathbb{S}^2$; note that A is defined modulo the area of $\mathbb{S}^2$, namely 4π, which is exactly the length of the fiber above the initial point of α. However, this geometry is not maximal. Indeed, it is well known that there is a diffeomorphism between $\widetilde{T}^1\mathbb{S}^2$ and $\mathbb{S}^3$ which sends the projection $\widetilde{T}^1\mathbb{S}^2 \to \mathbb{S}^2$ to the Hopf fibration. In addition, an immediate computation shows that the standard identification $\mathbb{S}^3 \cong \widetilde{T}^1\mathbb{S}^2$ sends the metric of $\mathbb{S}^3$ to the metric obtained by rescaling the metric of $\widetilde{T}^1\mathbb{S}^2$ by a factor of $\frac{1}{2}$ in the direction of the fibers (so that the fibers now have length 2π). As a consequence, the geometry of $\widetilde{T}^1\mathbb{S}^2 = \mathbb{S}^2 \mathbin{\tilde{\times}} \mathbb{S}^1$ is actually contained in the geometry of $\mathbb{S}^3$, and is not maximal.

We will later see that the four geometries considered in this section mostly occur for manifolds which admit fibrations of a certain type, called Seifert fibrations. For this reason, these are often called Seifert-type geometries.

2.3. *The* Sol *geometry*

Finally, there is a maximal geometry (X, G) where all stabilizers G_x are finite.

Topologically, X is just $\mathbb{R}^3$, but is endowed with the Riemannian metric m_0 which at (x, y, z) is $ds^2 = \mathrm{e}^{2z}\,dx^2 + \mathrm{e}^{-2z}\,dy^2 + dz^2$. (There is nothing special about the number e; it can easily be replaced by any number greater than 1 by rescaling in the z-direction.) This metric is respected by the group G of transformations of X of the form

$$(x, y, z) \mapsto \left(\varepsilon \mathrm{e}^{-c}x + a, \varepsilon' \mathrm{e}^{c}y + b, z + c\right) \quad \text{or}$$
$$\left(\varepsilon \mathrm{e}^{-c}y + a, \varepsilon' \mathrm{e}^{c}x + b, -z + c\right),$$

where a, b, $c \in \mathbb{R}$ and ε, $\varepsilon' = \pm 1$. Namely, G is generated by all horizontal translations, the reflections across the xz- and yz-coordinate planes, the vertical shifts $(x, y, z) \mapsto$

$(e^{-c}x, e^{c}y, z+c)$, $c \in \mathbb{R}$, and the flip $(x, y, z) \mapsto (y, x, -z)$. Note that the stabilizer of the origin consists of the eight maps $(x, y, z) \mapsto (\pm x, \pm y, z)$ and $(x, y, z) \mapsto (\pm y, \pm x, -z)$.

Let us show that the geometry (X, G) is maximal. As usual, it suffices to show that, for any G-invariant metric m, the isometry group of m is not larger than G. Looking at stabilizers, we immediately see that such a G-invariant metric m must be obtained by rescaling the original metric m_0 in the horizontal and vertical directions, namely that there exist constants λ, $\mu > 0$ such that the metric m corresponds to $ds^2 = \lambda\, e^{2z}\, dx^2 + \lambda\, e^{-2z}\, dy^2 + \mu\, dz^2$. Then, the sectional curvature of m is equal to $+1/\mu$ along any horizontal (for the identification $X = \mathbb{R}^3$) tangent plane, is equal to $-1/\mu$ along any vertical tangent plane, and is strictly between these two values along a plane which is neither horizontal nor vertical. It follows that any m-isometry φ must respect up to sign the vector field U which is vertical pointing upwards in $X = \mathbb{R}^3$ and has norm 1 for m. If we consider the covariant differentiation $v \mapsto \nabla_v U$ as an automorphism of each tangent space $T_{(x,y,z)}X = \mathbb{R}^3$, a straightforward computation shows that the direction of the x-axis is the eigenspace of a positive eigenvalue, and that the direction of the y-axis is the eigenspace of a negative eigenvalue. Therefore, the differential of an m-isometry φ must respect the three coordinate axes if φ sends U to U, and exchange the x- and y-axes if φ sends U to $-U$. It easily follows that there is an element φ' of G which has the same value and the same differential as φ at an arbitrary point $(x_0, y_0, x_0) \in X$, from which we conclude that $\varphi = \varphi' \in G$. This concludes the proof that the geometry (X, G) is maximal.

The group G is easily seen to be solvable, and is the only one with this property among the groups corresponding to the geometries we have encountered so far. For this reason, the geometry (X, G) is often called the Sol geometry.

At this point, we have described eight maximal 3-dimensional geometries. We will later see that, for each of these geometries, there is at least one manifold which admits a finite volume complete geometric structure corresponding to this geometry. Thurston showed that the list is complete, and that there is no other maximal 3-dimensional geometry with this property. We cannot give the details of the proof of this fact here, and refer to Scott [126, §5] for a discussion of this proof. However, it is probably worth mentioning that certain maximal geometries, such as the geometry of $X = \mathbb{R}^3$ endowed with the isometry group G of the metric $ds^2 = e^{2\lambda z}\, dx^2 + e^{-2\mu z}\, dy^2 + dz^2$ with $\lambda \neq \mu$ are excluded because no manifold admits a *finite volume* (G, X)-structure of this type.

2.4. *Topological obstructions to the existence of complete geometric structures*

We saw that every surface admits a complete geometric structure. In dimension 3, there are topological obstructions to the existence of a complete geometric structure on a given 3-manifold.

A simple observation restricts the geometric structures with which a non-orientable manifold M can be endowed. We saw that the geometries of $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ and $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ are *chiral*, in the sense that they admit no orientation-reversing isometries. In particular, in the atlas defining on M a geometric structure modelled over one of these two spaces, the changes of charts are orientation-preserving. It follows that M is orientable.

The same holds for a *complete* geometric structure modelled over $\mathbb{S}^3$. Indeed, if M is endowed with a complete geometric structure modelled over $\mathbb{S}^3$, any isometry between an open subset of the universal covering $\widetilde{M}$ of M and an open subset of $\mathbb{S}^3$ uniquely extends to an isometry between $\widetilde{M}$ and $\mathbb{S}^3$, by the result of Singer [131] which we already mentioned. As a consequence, M is isometric to a quotient $\mathbb{S}^3/\Gamma$, where $\Gamma \cong \pi_1(M)$ is a finite group acting freely and isometrically on $\mathbb{S}^3$. The Lefschetz Fixed Point Theorem, or inspection, shows that every orientation-reversing isometry of $\mathbb{S}^3$ must have fixed points. Therefore, Γ must respect the orientation of $\mathbb{S}^3$. This proves:

PROPOSITION 2.1. *If the 3-manifold M admits a geometric structure modelled over* $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ *or* $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$, *or a complete geometric structure modelled over* $\mathbb{S}^3$, *then M is orientable.*

The second restriction has to do with the fact that, for seven out of the eight 3-dimensional geometries, the model space X is homeomorphic to $\mathbb{R}^3$ or $\mathbb{S}^3$ and, as such, contain no essential 2-sphere. A 2-sphere S embedded in the 3-manifold M is *essential* if the closure of no component of $M - S$ is diffeomorphic to the 3-ball. By a theorem of Alexander [4], $\mathbb{R}^3$ and $\mathbb{S}^3$ contain no essential 2-sphere. Among the eight 3-dimensional geometries, the only model space that contain essential 2-spheres is therefore $\mathbb{S}^2 \times \mathbb{E}^1$.

An *essential projective plane* in the 3-manifold M is a surface P embedded in M which is diffeomorphic to the real projective plane $\mathbb{RP}^2$ and which is 2-sided, namely such that the normal bundle of P in M is trivial. Note that, if M contains a 1-sided projective plane, either M contains an essential 2-sphere, namely the boundary of a tubular neighborhood of P, or else M is diffeomorphic to the real projective 3-space $\mathbb{RP}^3$.

An easy covering theory argument shows that, if we have a covering $\widetilde{M} \to M$ of 3-manifolds and if $\widetilde{M}$ contains no essential 2-sphere or projective plane, then M also contains no essential 2-sphere or projective plane. The converse is actually true by a deeper result, the Equivariant Sphere Theorem of Meeks, Simon and Yau [88–90]. But we only need the result in the forward direction, which immediately shows:

THEOREM 2.2. *If the 3-manifold M admits a complete geometric structure modelled over* $\mathbb{E}^3$, $\mathbb{S}^3$, $\mathbb{H}^3$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ *or* Sol *(namely one of the eight 3-dimensional geometries except* $\mathbb{S}^2 \times \mathbb{E}^1$*), then M contains no essential 2-sphere or projective plane.*

More stringent restrictions occur for the Seifert-type geometries. Namely, the existence of such a geometry on a compact 3-manifold usually leads to a Seifert fibration on this manifold. But we first need to define Seifert fibrations. Seifert fibrations were introduced by Seifert[3] [128]; in addition to [128], useful references on Seifert fibrations include the books by Orlik [104] and Montesinos [97], as well as Scott's survey [126].

A *Seifert fibration* of the 3-manifold M is a decomposition of M into disjoint simple closed curves, called the *fibers* of the fibration, such that the following property holds: Every fiber has a neighborhood U which is diffeomorphic to the quotient of a solid torus $\mathbb{S}^1 \times \mathbb{B}^2$ by the free action of a finite group action respecting the product structure, in such

[3] The fibrations considered by Seifert had only orientation-preserving fibers, but the generalization given below is not intrinsically different.

a way that the fibers of the fibration correspond to the images of the circles $\{x\} \times \mathbb{B}^2$. Here, $\mathbb{B}^n$ denotes the closed unit ball in $\mathbb{R}^n$.

By inspection, the above fibered neighborhood U has to be of one of the following two types:

If U is orientable, then there is a diffeomorphism between U and $\mathbb{S}^1 \times \mathbb{B}^2$ for which, if we identify $\mathbb{S}^1$ and $\mathbb{B}^2$ to the unit circle and unit disk in $\mathbb{C}$, the fibers of U all have a parametrization of the form $z \mapsto (z^p, z_0 z^q)$, where z ranges over $\mathbb{S}^1$, $z_0 \in \mathbb{B}^2$ depends on the fiber, and the coprime integers $p > 0$ and q depend only on U and on its parametrization. Namely, the fibers wrap p times in the $\mathbb{S}^1$-direction and q times in the $\mathbb{B}^2$-direction, except the central fiber corresponding to $z_0 = 0$, which wraps only once in the $\mathbb{S}^1$-direction.

If U is non-orientable, it admits a diffeomorphism with $[0, 1] \times \mathbb{B}^2/\!\sim$, where $\sim$ identifies $\{1\} \times \mathbb{B}^2$ to $\{0\} \times \mathbb{B}^2$ by complex conjugation, and where the fibers correspond to the sets $[0, 1] \times \{z_0, \bar{z}_0\}$. Note that most fibers are orientation-preserving, with the exception of those corresponding to $z_0 \in \mathbb{R}$.

A fiber is *generic* if it admits an orientable fibered neighborhood U as above with $p = 1$ and $q = 0$, namely if the fibration is a locally trivial bundle near this fiber; otherwise, the fiber is *exceptional*. Orientation-preserving exceptional fibers are isolated. Orientation-reversing exceptional fibers form a 2-dimensional submanifold of M whose components are open annuli, tori and Klein bottles (since they are locally trivial circle bundles).

Now, consider the space Σ of fibers of a Seifert fibration of the 3-manifold M. Near a fiber f which is orientation-preserving in M, a fibered solid torus neighborhood U of f in M provides a neighborhood of the point $f \in \Sigma$ in Σ which is homeomorphic to a quotient $\mathbb{B}^2/\mathbb{Z}_p$, where the cyclic group $\mathbb{Z}_p$ (possibly with $p = 1$) acts on $\mathbb{B}^2$ by rotation; note that this quotient $\mathbb{B}^2/\mathbb{Z}_p$ is homeomorphic to a disk centered at f. Similarly, the point of Σ corresponding to a fiber f which is orientation-reversing in M has a neighborhood homeomorphic to $\mathbb{B}^2/\mathbb{Z}_2$, where $\mathbb{Z}_2$ acts on $\mathbb{B}^2$ by complex conjugation; note that $\mathbb{B}^2/\mathbb{Z}_2$ is in this case homeomorphic to a half-disk with f on its boundary. As a consequence, Σ is a topological surface with boundary, where the boundary points correspond to orientation-reversing fibers.

When we consider Σ only as a topological surface, we unfortunately loose a lot of information regarding the differentiable structure of M. For instance, we can endow the surface Σ with a differentiable structure for which the natural projection $p: M \to \Sigma$ is differentiable but, if the fibration has at least one exceptional fiber, there is no differentiable structure on Σ for which $M \to \Sigma$ is a submersion, in contrast to what is usually expected of a fibration. For this reason, it is much better to consider the natural orbifold structure of Σ. This leads us to digress into a brief discussion of orbifolds.

Orbifolds were first introduced in the 1950s by Satake [121,122] under the name of V-manifolds, and later rediscovered and popularized by Thurston [138] under the name of orbifolds. In addition to these references, some basic facts about orbifolds can also be found in Montesinos [97] or Bonahon and Siebenmann [20].

Roughly speaking, an orbifold is a topological space endowed with an atlas which locally models it over quotients of manifolds by properly discontinuous group actions. More precisely, let a (differentiable) *folding map* be a continuous map $f: \widetilde{U} \to U$ from a differentiable manifold $\widetilde{U}$ to a topological space U such that the *folding group* G_f, defined as the group of diffeomorphisms g of $\widetilde{U}$ such that $f \circ g = f$, acts properly discontinuously[4] on $\widetilde{U}$ and such that the induced map $\widetilde{U}/G_f \to U$ is a homeomorphism.

A (differentiable) *orbifold* is defined as a metrizable topological space O endowed with an atlas of *folding charts* $f_i : \widetilde{U}_i \to U_i$, $i \in I$, where the U_i form an open covering of O and where the f_i are compatible in the following sense: For every $x_i \in \widetilde{U}_i$ and $x_j \in \widetilde{U}_j$ with $f_i(x_i) = f_j(x_j)$, there is a diffeomorphism ψ_{ij} from an open neighborhood $\widetilde{V}_i$ of x_i in $\widetilde{U}_i$ to an open neighborhood $\widetilde{V}_j$ of x_j in $\widetilde{U}_j$ such that $f_j \circ \psi_{ij} = f_i$ over $\widetilde{V}_i$. More formally, an orbifold is a topological space O with a maximal atlas of compatible folding charts as above. Note that it is always possible to restrict the folding charts so as to obtain an atlas where all folding charts have finite folding group. So, the definition of orbifolds would be unchanged if we restricted attention to folding charts with finite folding groups, which is what many authors do. To alleviate the notation, we will often use the same symbol to represent an orbifold and its underlying topological space; in theory, this is somewhat dangerous since the structure of an orbifold involves much more data than its underlying topological space, but we will try to make sure that the context clearly identifies the interpretation which has to be used. When we really need to emphasize the distinction, we will denote by $|O|$ the topological space underlying the orbifold O.

A typical example of orbifold is provided by the properly discontinuous action of a group Γ over a manifold M. Then, the quotient map $M \to M/\Gamma$ is a folding chart, and defines an orbifold structure on M/Γ. An orbifold obtained in this way is said to be *uniformizable*. Although many (and perhaps most) important orbifolds are uniformizable, there exists orbifolds which are not; we will encounter some non-uniformizable 2-orbifolds in Proposition 2.6. In any case, it is always useful to keep the example of uniformizable orbifolds in mind, since any orbifold is *locally* of this type.

An *orbifold covering map* between two orbifolds is a continuous open map $\varphi : O \to O'$ between their underlying topological spaces such that, if $\{f_i : \widetilde{U}_i \to U_i;\ i \in I\}$ is the maximal atlas defining the first orbifold, $\{\varphi \circ f_i : \widetilde{U}_i \to \varphi(U_i);\ i \in I\}$ is an atlas defining the second orbifold. If, in addition, the map $\varphi : O \to O'$ is a homeomorphism, then φ^{-1} is also an orbifold covering map, and this defines an *isomorphism* between the two orbifolds.

If the group Γ acts properly discontinuously on M and if Γ' is a subgroup of Γ, the canonical map $M/\Gamma' \to M/\Gamma$ is an orbifold covering map. By definition, an orbifold covering map is always *locally* of this type.

For every point x of the topological space O underlying an orbifold, for every folding chart $f_i : \widetilde{U}_i \to U_i$ of the orbifold atlas with $x \in U_i$ and for every $\widetilde{x} \in f_i^{-1}(x)$, it immediately follows from definitions that the action on the tangent space $T_{\widetilde{x}}\widetilde{U}_i$ of the (finite) stabilizer of $\widetilde{x}$ in the folding group G_{f_i} depends only on x, up to conjugation. The corresponding finite linear group G_x, well defined up to linear conjugation, is the *isotropy group* of x. The point x is *regular* if the isotropy group G_x is trivial, and *singular* otherwise. For instance, for the uniformizable orbifold M/Γ arising from a properly discontinuous action of a group Γ on a manifold M, the set of singular points of M/Γ is exactly the image of the union of the fixed point sets of the non-trivial elements of Γ.

By straightforward generalization of the case of manifolds, we can define geometric structures on orbifolds. Namely, an orbifold admits an (X, G)-*structure* if its maximal

[4] Recall that a group G acts *properly discontinuously* on a locally compact space X if every $x \in X$ admits a neighborhood V such that $\{g \in G;\ V \cap gV \neq \emptyset\}$ is finite.

orbifold atlas contains an atlas $\{f_i : \widetilde{U}_i \to U_i;\ i \in I\}$ where the $\widetilde{U}_i$ are open subsets of X, the folding groups G_{f_i} consist of restrictions to $\widetilde{U}_i$ of elements of G, and the change of charts ψ_{ij} are also restrictions of elements of G. If X is endowed with a G-invariant Riemannian metric, this metric induces a Riemannian metric on the set of regular points of the orbifold, and therefore a metric space structure on the topological space underlying the orbifold, by defining the distance from x to y as the infimum of the lengths of those paths which go from x to y and which consist only of regular points, with the possible exception of x and y. By definition, the corresponding geometric structure is *complete* if this underlying metric space is complete.

If an orbifold O admits a complete (X, G)-structure, consider a folding chart $f : \widetilde{U} \to U$ of the atlas defining this structure. By definition, $\widetilde{U}$ is an open subset of X and the folding group G_f is a subgroup of G. Then, the argument of Singer [131] immediately extends to give a global folding chart $X \to O$, whose folding group Γ is contained in G. In other words the orbifold O is isomorphic to the orbifold X/Γ, quotient of X by a subgroup Γ of G acting properly discontinuously on X. (See also Thurston [138, Chapter 3] or Benedetti and Petronio [11, Section B.1].) This proves:

LEMMA 2.3. *If an orbifold admits a complete geometric structure modelled over (X, G), then it is isomorphic to the orbifold X/Γ quotient of X by a subgroup Γ of G acting properly discontinuously on X, and this by an isomorphism respecting geometric structures. In particular, the orbifold is uniformizable.*

In the case of the base space Σ of a Seifert fibration of the 3-manifold M, any small 2-dimensional submanifold $\widetilde{U}$ of M which is transverse to the leaves projects to an open subset U of Σ, and the local models for the Seifert fibration show that the restriction of the projection p to $\widetilde{U} \to U$ locally is a folding chart. It is immediate that these folding charts are compatible, and therefore define an orbifold structure on Σ. This 2-dimensional orbifold is the *base orbifold* of the Seifert fibration. Note that the singular points of this orbifold are exactly those corresponding to exceptional fibers of the Seifert fibration; the isotropy group of such a singular point is cyclic, acting by rotations on $\mathbb{R}^2$, when the singular point corresponds to an orientation-preserving exceptional fiber, and is $\mathbb{Z}_2$ acting by reflection when it corresponds to an orientation-reversing exceptional fiber.

Up to orbifold isomorphism, this base 2-orbifold is completely determined by: the topological type of the topological surface Σ with boundary underlying the orbifold; the discrete subset S of Σ consisting of those singular points where the isotropy group is a finite rotation group; the assignment of this isotropy group G_x to each $x \in S$. This easily follows from local considerations near the singular set.

Seifert fibrations were classified by Seifert [128] in the 1930s. Namely, given two 3-manifolds M and M' each endowed with a Seifert fibration, he introduced invariants which enabled him to decide whether there exists a diffeomorphism between M and M' sending fibration to fibration. As indicated earlier, Seifert was only considering fibrations where the fibers are orientation-preserving, but his work straightforwardly extends to the case where we allow orientation-reversing fibers. We now discuss Seifert's classification, using a reformulation proposed by Thurston. The details of this classification can be found in Seifert [128], Orlik [104], Scott [126], Montesinos [97], Bonahon and Siebenmann [20].

We first consider the classification of Seifert fibrations of oriented 3-manifolds M.

In this case, there are no orientation-reversing fibers, so that the topological space underlying the base 2-orbifold Σ is a surface without boundary. The first invariant of the Seifert fibration is the orbifold isomorphism type of Σ.

Then, there is an invariant $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ associated to each exceptional fiber f as follows: Let $U \cong \mathbb{S}^1 \times \mathbb{B}^2$ be a fibered neighborhood of f where the fibers have a parametrization of the form $z \mapsto (z^p, z_0 z^q)$, $z \in \mathbb{S}^1$, where f is the central fiber corresponding to $z_0 = 0$, and where the identification $U \cong \mathbb{S}^1 \times \mathbb{B}^2$ is consistent with the orientation of M and with the standard orientation of $\mathbb{S}^1$ and $\mathbb{B}^2$. Then $\alpha = p$ and $\beta \in \mathbb{Z}$ is such that $\beta q \equiv 1 \bmod p$. (In particular, the data of $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ is equivalent to that of $q/p \in \mathbb{Q}/\mathbb{Z}$, but it turns out to be slightly more convenient.) Note that, if we consider f as a point of Σ, the isotropy group of f in the base orbifold is $\mathbb{Z}_\alpha$, acting by rotations.

Finally, when M is compact, there is a globally defined Euler number $e_0 \in \mathbb{Q}$. This invariant has the property that its mod $\mathbb{Z}$ reduction is equal to the sum of the invariants $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ associated to all the exceptional fibers of the fibration. When the Seifert fibration is a locally trivial bundle where the fibers can be coherently oriented, this bundle has an Euler class defined in the cohomology group $H^2(\Sigma; \mathbb{Z})$; the orientations of M and of the fibers determine an orientation of Σ, which itself determines an isomorphism $H^2(\Sigma; \mathbb{Z}) \cong \mathbb{Z}$; then, e_0 is the integer corresponding to the Euler class through this isomorphism; note that reversing the orientation of the fibers multiplies the Euler class and the isomorphism $H^2(\Sigma; \mathbb{Z}) \cong \mathbb{Z}$ by -1, so that e_0 is unchanged. Also, this Euler number is well behaved with respect to coverings: If we have a finite covering $M' \to M$ of oriented manifolds and if M is endowed with a Seifert fibration of Euler number e_0, this fibration pulls back to a Seifert fibration of M' whose Euler number is $e_0 n^2/p$, where p is the degree of the covering and where n is the number of components of the preimage in M' of a generic fiber of M. With the fact that $e_0 = 1$ for every Seifert fibration of $\mathbb{S}^3$, these properties can actually be used to explicitly define e_0: Indeed, an easy exercise, based on the choice of suitable orbifold coverings of Σ, shows that, for every Seifert fibration of M, there is a finite covering of M where this Seifert fibration pulls back to a locally trivial bundle or to a Seifert fibration of $\mathbb{S}^3$. A more explicit definition of e_0 can be found in Neumann and Raymond [101], Montesinos [97], or Bonahon and Siebenmann [20].

When M is not compact, e_0 is not defined.

Seifert's classification of Seifert fibrations of oriented 3-manifolds can be rephrased as follows.

THEOREM 2.4 (Oriented classification of Seifert fibrations). *Consider two oriented 3-manifolds M and M', each endowed with a Seifert fibration. Then, there is an orientation-preserving diffeomorphism $M \to M'$ sending fiber to fiber if and only if there is an isomorphism between their base orbifolds which sends each singular point to a singular point with the same invariant $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ and if, when the manifolds are compact, the two Seifert fibrations have the same Euler number $e_0 \in \mathbb{Q}$. Conversely, if Σ is a 2-orbifold where all isotropy groups are cyclic acting by rotation, if we assign to each singular point of Σ with isotropy group $\mathbb{Z}_\alpha$ an element $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ with β coprime to α, and if, when Σ is compact, we pick a rational number $e_0 \in \mathbb{Q}$ whose* mod $\mathbb{Z}$ *reduction is equal to the sum of the β/α, there is a Seifert fibration of an oriented 3-manifold M which realizes this data.*

Note that reversing the orientation of the manifold M reverses the sign of the invariants $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ associated to their exceptional fibers and, if applicable, of the Euler number $e_0 \in \mathbb{Q}$. Therefore, Theorem 2.4 also provides the *unoriented* classification of Seifert fibrations of orientable 3-manifolds.

The classification of Seifert fibrations of non-orientable manifolds is somewhat harder to state, but it essentially follows the lines of the oriented classification. We can summarize it by saying that such a Seifert fibration is characterized by the following data: the base orbifold Σ; invariants $\beta/\alpha \in \mathbb{Q}/\mathbb{Z}$ associated to the orientation-preserving exceptional fibers; orientability data for the locally trivial bundle obtained by removing the exceptional fibers; when the manifold is compact, a global obstruction in $\mathbb{Z}_2$. However these data tend to be interdependent. We refer to Seifert [128], Scott [126], Montesinos [97], Bonahon and Siebenmann [20] for precise statements.

We can now state the restrictions which a Seifert-type geometry imposes on a 3-manifold. Complete proofs and details can be found in Scott [126].

THEOREM 2.5. *If the manifold M admits a complete geometric structure modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ or $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$, then one of the following occurs:*

(i) *The foliation of M by the $\mathbb{E}^1$ factors is a Seifert fibration. In this case, all generic fibers of the Seifert fibration have the same length l, and the metric of the $\mathbb{S}^2$, $\mathbb{H}^2$ or $\mathbb{E}^2$ factors projects to a complete spherical, hyperbolic or Euclidean geometric structure on the base orbifold Σ of this fibration. If M is compact, orientable, and oriented so that the charts of its geometric structure are orientation-preserving, then the Euler number $e_0 \in \mathbb{Q}$ is equal to 0 for the product geometry of $\mathbb{S}^2 \times \mathbb{E}^1$ and $\mathbb{H}^2 \times \mathbb{E}^1$, and is negative for the twisted geometries of $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ and $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$. In addition, when e_0 is defined and non-zero, the length l of the generic fibers is equal to $-e_0\,\mathrm{area}(\Sigma)$, where the area is that of the geometric structure induced on the base 2-orbifold Σ.*

(ii) *The foliation of M by the $\mathbb{E}^1$ factors is a (locally trivial) $\mathbb{E}^1$-bundle over a surface Σ. In this case, the metric of the $\mathbb{S}^2$, $\mathbb{H}^2$ or $\mathbb{E}^2$ factors projects to a spherical, hyperbolic or Euclidean complete geometric structure on Σ.*

(iii) *At most two leaves of the foliation by the $\mathbb{E}^1$ factors are closed subsets of M. In this case, M is diffeomorphic to one of eleven manifolds: the two $\mathbb{S}^2$-bundles over $\mathbb{S}^1$, the connected sum $\mathbb{RP}^3 \# \mathbb{RP}^3$ of two copies of the real projective 3-space $\mathbb{RP}^3$, the product $\mathbb{RP}^2 \times \mathbb{S}^1$, the two $\mathbb{E}^2$-bundles over $\mathbb{S}^1$, the two $\mathbb{E}^1$-bundles over the 2-torus, or the three $\mathbb{E}^1$-bundles over the Klein bottle.*

If, in addition, the geometric structure of M has finite volume, then case (ii) *and the non-compact manifolds of case* (iii) *cannot occur. In case* (i), *the geometric structure induced on the base 2-orbifold Σ has finite area. In case* (iii), *the geometric structure of M is necessarily modelled over $\mathbb{S}^2 \times \mathbb{E}^1$.*

The conclusions of case (i) will be more useful if we can specify which 2-orbifolds admit complete spherical, Euclidean, or hyperbolic structures.

For this, it is convenient to introduce the Euler characteristic of a compact orbifold O. The proof that every differentiable manifold admits a triangulation immediately extends to show that every differentiable orbifold admits a triangulation. A *triangulation* of the orbifold O is a decomposition of the topological space underlying O into subsets of O called *orbifold simplices*, such that each point x of this underlying space has a neighborhood U which is a union of orbifold simplices, which is contained in the image of some folding chart $f_i : \widetilde{U}_i \to U_i$ of the orbifold atlas of O, and such that the decomposition of U into orbifold simplices lifts to a triangulation of $f_i^{-1}(U) \subset \widetilde{U}_i$ which is invariant under the action of the folding group G_{f_i}. In addition, we insist that the isotropy group is constant on the interior of each orbifold simplex. Then, the *orbifold Euler characteristic* of the compact orbifold O is the sum

$$\chi_{\mathrm{orb}}(O) = \sum_{\sigma} (-1)^{\dim \sigma} \frac{1}{|G_\sigma|} \in \mathbb{Q},$$

where the sum is over all orbifold simplices σ of the triangulation σ, and where $|G_\sigma|$ is the cardinal of the isotropy group of an arbitrary point of the interior of σ. Standard proofs that the Euler characteristic of a manifold is independent of the triangulation automatically extend to orbifolds. Note that the orbifold characteristic $\chi_{\mathrm{orb}}(O)$ is a rational number, and should not be confused with the usual Euler characteristic $\chi(|O|)$ of the topological space $|O|$ underlying the orbifold O.

We similarly define the orbifold Euler characteristic of an orbifold O of *finite type*, namely an orbifold which is isomorphic to the interior of a compact orbifold $\bar{O}$ with boundary (where orbifolds with boundary are defined by replacing manifolds by manifolds with boundary in the definition of folding maps). In this case, $\chi_{\mathrm{orb}}(O) = \chi_{\mathrm{orb}}(\bar{O})$.

A fundamental property of this orbifold Euler characteristic is that it is well behaved with respect to orbifold coverings: If there is an orbifold covering $O \to O'$ of degree n, namely such that the pre-image of a regular point of O' consists of n regular points of O, then $\chi_{\mathrm{orb}}(O) = n\chi_{\mathrm{orb}}(O')$. Also, note that the orbifold Euler characteristic $\chi_{\mathrm{orb}}(O)$ coincides with the usual Euler characteristic when the orbifold O is a manifold, namely when all the isotropy groups are trivial.

When Σ is a finite type 2-orbifold of the type occurring as base orbifolds of Seifert fibrations, namely where all isotropy groups are cyclic acting by rotations or $\mathbb{Z}_2$ acting by reflection, it is immediate from definitions that $\chi_{\mathrm{orb}}(\Sigma)$ is the sum of the usual Euler characteristic of its underlying space $|\Sigma|$, of $-\frac{1}{2}$ times the number of non-compact components of the set of reflection points of Σ, and of $\sum_{i=1}^{k}(1/\alpha_i - 1)$ where Σ has exactly k isolated singular points and the isotropy group of the ith singular point is $\mathbb{Z}_{\alpha_i}$ acting by rotation.

PROPOSITION 2.6 (Geometrization of 2-orbifolds). *Let Σ be a 2-orbifold of the type occurring as base orbifolds of Seifert fibrations, namely where all isotropy groups are cyclic acting by rotations or $\mathbb{Z}_2$ acting by reflection. Then*:

(i) *If Σ is compact, it admits a hyperbolic structure if and only if its orbifold Euler characteristic $\chi_{\mathrm{orb}}(\Sigma)$ is negative.*

(ii) *If Σ is compact, it admits a Euclidean structure if and only if $\chi_{\mathrm{orb}}(\Sigma)$ is equal to* 0.

(iii) *If Σ is compact, it admits a spherical structure only if $\chi_{\mathrm{orb}}(\Sigma)$ is positive. Conversely, if $\chi_{\mathrm{orb}}(\Sigma) > 0$, either Σ admits a spherical structure, or Σ has underlying* *topological space a 2-sphere and exactly one singular point, or Σ has underlying topological space a 2-sphere and has exactly two singular points, of respective isotropy groups $\mathbb{Z}_p$ and $\mathbb{Z}_q$ with $p \neq q$.*

(iv) *If Σ is non-compact, it always admits a complete hyperbolic structure.*

(v) *If Σ is non-compact, it admits a complete Euclidean structure if and only if it falls into one of the following eight categories: the topological space underlying Σ is a plane and Σ has at most one singular point; the topological space underlying Σ is a half-plane, and the set of singular points of Σ forms a line; the topological space underlying Σ is a plane and Σ has exactly two singular points, with isotropy group $\mathbb{Z}_2$ acting by rotations; the topological space underlying Σ is a half-plane, and the singular set of Σ consists of one line of reflection points and of a single point with isotropy group $\mathbb{Z}_2$ acting by rotation; the topological space underlying Σ is an open annulus, with no singular point; the topological space underlying Σ is a semi-open annulus, and the singular set consists of a circle of reflection points; the topological space underlying Σ is a square with two opposite sides removed, and the singular set consists of two lines; the topological space underlying Σ is an open Möbius strip, with no singular point.*

(vi) *If M is non-compact, it cannot admit a complete spherical structure.*

Here, a half-plane is the closure of one component of the complement of a line in $\mathbb{E}^2$. A *closed annulus* is a 2-manifold diffeomorphic to $\mathbb{S}^1 \times [0, 1]$. If we remove 1 or 2 boundary components from a compact annulus, one gets a *semi-open* or *open* annulus. The terminology is similar for closed and open Möbius strips.

In case (iii) of Proposition 2.6, we encounter two exceptional types of 2-orbifolds: those where Σ has underlying topological space a 2-sphere and exactly one singular point; and those where Σ has underlying topological space a 2-sphere and has exactly two singular points, of respective isotropy groups $\mathbb{Z}_p$ and $\mathbb{Z}_q$ with $p \neq q$. In either case, an easy covering argument on the complement of the singular set shows that these orbifolds are not uniformizable, namely that they are isomorphic to no orbifold M/Γ, where the group Γ acts properly discontinuously on the 2-manifold M. In particular, by Lemma 2.3, these orbifolds admit no geometric structure.

Proposition 2.6 can be straightforwardly generalized to include all 2-orbifolds, allowing dihedral isotropy groups. The list of exceptions is just a little longer. Our restriction to cyclic isotropy groups is only for the convenience of the exposition, since the only 2-orbifolds which we will encounter in this chapter mostly arise as base orbifolds of Seifert fibrations.

The proof of Proposition 2.6 is fairly elementary. The necessary conditions on the Euler characteristic in cases (i), (ii) and (iii) follow from an immediate extension of the Gauss–Bonnet formula to 2-orbifolds. The existence part can be proved by methods similar to those used in Section 1.1 to construct geometric structures on surfaces. The analytic method generalizes to orbifolds without any major problems (Hint: First construct a 'universal orbifold covering' $\widetilde{\Sigma} \to \Sigma$ and show that $\widetilde{\Sigma}$ has no singular point unless Σ is one of the exceptional orbifolds of case (ii); see also [85]), and it is interesting to see why it fails to provide a geometric structure on the non-uniformizable 2-orbifolds. For orbifolds of finite type, the geometric cut-and-paste method is probably easier. In particular, this is the method used in [102] where a complete proof of Proposition 2.6 for finite type 2-orbifolds is given.

After this digression on geometric 2-orbifolds, we now return to 3-manifolds.

A surprising fact is that a geometry modelled over $\mathbb{E}^3$ also leads to a Seifert fibration. Again, a detailed proof can be found in Scott [126].

THEOREM 2.7. *If the manifold M admits a complete geometric structure modelled over the Euclidean space $\mathbb{E}^3$, then the maximal atlas defining the geometric structure contains an atlas modelling M over $\mathbb{E}^2 \times \mathbb{E}^1 = \mathbb{E}^3$, where all changes of charts respect the splitting of $\mathbb{E}^2 \times \mathbb{E}^1$, and such that at least one of the following occurs:*

(i) *The foliation of M by the $\mathbb{E}^1$ factors defines a Seifert fibration of M; the metric of the $\mathbb{E}^2$ factor induces a euclidean structure on the base orbifold of the Seifert fibration and, if defined, the Euler number e_0 of the fibration is equal to 0.*

(ii) *The foliation of M by the $\mathbb{E}^2$ factors endows M with the structure of a locally trivial bundle over the circle $\mathbb{S}^1$ with fiber the plane $\mathbb{E}^2$; topologically, there are exactly two such bundles.*

(iii) *M is the Euclidean space $\mathbb{E}^3$.*

Similarly, a geometric structure modelled over $\mathbb{S}^3$ leads to the existence of a Seifert fibration. As usual, we refer to [126] for a proof.

THEOREM 2.8. *If the manifold M admits a complete geometric structure modelled over the sphere $\mathbb{S}^3$, then the maximal atlas defining this geometric structure contains an oriented atlas modelling M over $\mathbb{S}^2 \tilde{\times} \mathbb{S}^1 = \mathbb{S}^3$, where all changes of charts respect the splitting of $\mathbb{S}^2 \tilde{\times} \mathbb{S}^1$. In addition, the foliation of M by the $\mathbb{S}^1$ factors defines a Seifert fibration; the metric of $\mathbb{S}^2$ defines a spherical structure on the base of this fibration and the Euler number e_0 of the fibration is strictly positive. In addition, the fundamental group of M is finite.*

We should probably emphasize the coincidental nature of Theorems 2.7 and 2.8. For instance, similar statements for 3-orbifolds are false. There are compact 3-orbifolds which admit euclidean or spherical structures but which admit no fibration; see Section 3.6.

We now turn to hyperbolic structures.

Let $\mathbb{T}^2$ denote the 2-torus $\mathbb{S}^1 \times \mathbb{S}^1$. A *singular torus* in the manifold M is a continuous map $\mathbb{T}^2 \to M$; it is *incompressible* if the induced homomorphism $\pi_1(\mathbb{T}^2) \to \pi_1(M)$ is injective. Let an *end* of the 3-manifold M be the image of a proper embedding $S \times [0, \infty[\to M$ where S is a compact surface without boundary. Recall that a map is *proper* when the pre-image of every compact set is compact. The reader should beware that what we call here an end for the sake of simplicity is usually called a tame end neighborhood of M. A singular torus $\mathbb{T}^2 \to M$ is *essential* if it is incompressible and if it cannot be homotoped to a singular torus with image in an end of M.

Note that, if the singular torus $\mathbb{T}^2 \to M$ is incompressible and can be homotoped into an end $U \cong S \times [0, \infty[$, then the fundamental group of U must contain a subgroup isomorphic to $\mathbb{Z}^2$, and S therefore is a 2-torus or a Klein bottle. Therefore, only the *toric* ends $U \cong S \times [0, \infty[$, where S is a torus or a Klein bottle, are relevant here.

THEOREM 2.9. *If the* 3*-manifold* M *admit a complete geometric structure modelled over the hyperbolic space* $\mathbb{H}^3$, *then* M *contains no essential* 2*-sphere, projective plane or singular torus. If, in addition, the hyperbolic structure of* M *has finite volume, then* M *is the union of a compact manifold with boundary and of finitely many toric ends.*

It should be observed that these topological restrictions to the existence of a hyperbolic structure are the weakest ones among those encountered in this section. This is consistent with what we already observed in Section 1.1 for the dimension 2, where all but finitely surfaces admitted a complete hyperbolic structure.

In most cases, if M admits a Seifert fibration with base orbifold Σ, then M contains an essential singular 2-torus. For instance, suppose that Σ contains a non-separating simple closed curve α which avoids the singular set, or a separating simple closed curve α avoiding the singular set such that each component of $\Sigma - \alpha$ topologically is neither a disk with 0 or 1 singular point, nor an open annulus with no singular point. Then, the fibers over α form an embedded 2-torus or Klein bottle in M, and it can be shown that the orientation cover of this surface gives an essential singular 2-torus in M. When this type of technique does not provide an essential singular 2-torus, other arguments show that the fundamental group of M is finite, or that M contains an essential 2-sphere or projective plane, all properties preventing it from admitting a complete hyperbolic structure. This yields the following property, whose proof can be found in [126, Chapter 5].

THEOREM 2.10. *Let the* 3*-manifold* M *admit a Seifert fibration with base* 2*-orbifold* Σ. *Then, it admits no complete hyperbolic structure, unless* Σ *is non-compact and admits a Euclidean structure (compare case* (v) *of Proposition* 2.6). *In addition,* M *admits no finite volume complete hyperbolic structure.*

There consequently is almost no overlap between those 3-manifolds which admit a complete Seifert-type geometric structure and those which admit a complete hyperbolic geometric structure.

Finally, we now discuss the Sol geometry. We first exhibit some 3-manifolds which admit such a geometry.

Let A be an element of $\mathrm{GL}_2(\mathbb{Z})$, namely a 2×2 matrix with integer entries and determinant ± 1. The matrix A defines a linear automorphism of $\mathbb{R}^2$ which respects the lattice $\mathbb{Z}^2$, and therefore induces a linear diffeomorphism φ of the 2-torus $\mathbb{T}^2 \cong \mathbb{R}^2/\mathbb{Z}^2$. Assume in addition that the eigenvalues of A are real and distinct; this is always the case when the determinant of A is -1 and, when the determinant is $+1$, occurs exactly when the trace of A has absolute value greater than 2. Since their product is equal to ± 1, we can write these eigenvalues as $\lambda_1 = \pm e^{-t}$, $\lambda_2 = \pm e^{t}$ with $t > 0$. By definition, this property of eigenvalues means that the linear diffeomorphism φ induced by A is an *Anosov linear diffeomorphism* of $\mathbb{T}^2$.

Choose a linear isomorphism $L : \mathbb{R}^2 \to \mathbb{R}^2 \times \{0\} \subset \mathbb{R}^3 = \mathrm{Sol}$ which sends the λ_1-eigenspace to the x-axis and the λ_2 to the y-axis. Consider the group Γ_0 of isometries of Sol which consists of all horizontal translations by elements of $L(\mathbb{Z}^2) \subset \mathbb{R}^2 \times \{0\}$, and let Γ be generated by Γ_0 and by the isometry $T : (x, y, z) \mapsto (\lambda_1 x, \lambda_2 y, z + t) = (\pm e^{-t} x, \pm e^{t} y, z + t)$. It is fairly immediate that Γ acts freely and properly discontinuously on Sol, and we can consider the quotient manifold $M = \mathrm{Sol}/\Gamma$, with the geometric structure induced by the geometric structure of Sol. Using the map L, we see that M is diffeomorphic to the *mapping torus* of the linear diffeomorphism φ, defines as the identification space $\mathbb{T}^2 \times [0, 1]/\sim$, where $\sim$ identified each point $(x, 0)$ to $(\varphi(x), 1)$.

In this way, we can put a geometric structure modelled over Sol on the mapping tori of all Anosov linear diffeomorphisms of the 2-torus.

This construction actually provides all non-trivial examples of 3-manifolds with a complete geometric structure modelled over Sol, as shown by the following result.

THEOREM 2.11. *If the 3-manifold M admits a complete geometric structure modelled over* Sol, *then one of the following occurs*:

(i) *The foliation of* $\mathrm{Sol} = \mathbb{R}^3$ *by horizontal planes induces a foliation of M whose leaves are 2-tori; this foliation actually defines on M the structure of a locally trivial bundle over $\mathbb{S}^1$ with fiber the 2-torus $\mathbb{T}^2$, and M is diffeomorphic to the mapping torus of an Anosov linear diffeomorphism of $\mathbb{T}^2$.*

(ii) *There is a geometry-preserving diffeomorphism between M and the quotient of* Sol *by a group of isometries which respects a horizontal plane; in particular, M is a line bundle over a 2-torus, a Klein bottle, an open annulus, an open Möbius strip or a plane.*

As usual, we refer to [126] for a proof of Theorem 2.11. Note that the list of 3-manifolds occurring in this context is extremely restricted.

2.5. *Geometric structures with totally geodesic boundary*

We now turn to geometric structures on manifolds with boundary. If we want to obtain any uniqueness properties for such geometric structures, we clearly have to impose some type of rigidity condition on the boundary. A natural condition is to require the boundary to be totally geodesic. Recall that a submanifold N of a Riemannian manifold M is *totally geodesic* if, locally, any small geodesic arc of M that joins two points of N is completely contained in N.

Let M be a manifold with boundary ∂M. Thicken M by gluing along ∂M a small collar $\partial M \times [0, 1[$, to obtain a manifold M^+ without boundary. By definition, a *geometric structure with totally geodesic boundary* on M is the restriction to M of a locally homogeneous Riemannian metric on M^+ for which the boundary ∂M is totally geodesic. Such a geometric structure is *complete* if the metric space structure induced on M is complete.

Among the eight 3-dimensional geometries, the two twisted geometries $\mathbb{H}^2 \tilde{\times} \mathbb{E}^1$ and $\mathbb{E}^2 \tilde{\times} \mathbb{E}^1$ locally contain no totally geodesic surface; this can easily be checked from the explicit expression of their metric given in Section 2.2.

All other six 3-dimensional geometries contain totally geodesic surfaces. For the isotropic geometries $\mathbb{S}^3$, $\mathbb{E}$, $\mathbb{H}^3$, there is such a totally geodesic surface passing through each point x and tangent to any plane in the tangent space at x. In the product geometries $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{S}^2 \times \mathbb{E}^1$, all totally geodesic surfaces locally are of the form $g \times \mathbb{E}^1$, $\mathbb{H}^2 \times \{x\}$ or

$\mathbb{S}^2 \times \{x\}$, where g is a geodesic of $\mathbb{H}^2$ or $\mathbb{S}^2$ and $x \in \mathbb{E}^1$. Finally, every totally geodesic surface of Sol is contained in a plane parallel to the xz- or yz-plane, in the model described in Section 2.3. In particular, these geometric models admit at most two complete totally geodesic surfaces, up to isometry. Therefore, for each geometry, there is at most two possible local models for a boundary point.

Note that, in each of the six model spaces $\mathbb{S}^3$, $\mathbb{E}^3$, $\mathbb{H}^3$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{S}^2 \times \mathbb{E}^1$, Sol admitting totally geodesic surfaces, there is an isometric involution of the model space which acts as a reflection across this surface. Conversely, it is easy to see that the fixed point set of an isometric involution is always totally geodesic. This observation provides a convenient way to analyze geometric structures with totally geodesic boundary, as follows.

Given a manifold M with boundary ∂M, let its *double* be the manifold DM without boundary obtained from the disjoint union of two copies M_1 and M_2 of M by gluing together the boundaries ∂M_1 and ∂M_2 through the natural identification maps $\partial M_1 \cong \partial M \cong \partial M_2$. This double DM comes equipped with a natural involution τ which exchanges the images of M_1 and M_2 and which fixes the image of the boundary ∂M. The differentiable structure of M gives a natural differentiable structure on DM for which the involution τ is differentiable. It then immediately follows from the above observations that the data of a geometric structure with totally geodesic boundary on M is equivalent to the data of a τ-invariant geometric structure on the double DM.

Now, the restrictions to the existence of a complete geometric structure on DM given in Section 2.4 easily translate to restrictions to the existence of complete geometric structures with totally geodesic boundary on M.

Let a *compression disk* for the boundary ∂M in M be a 2-dimensional submanifold D (with $D \cap \partial M = \partial D$) of M such that D is diffeomorphic to the disk (= 2-ball) $\mathbb{B}^2$. A compression disk is *essential* if its boundary ∂D does not bound a disk in the boundary ∂M. Similarly, a *singular compression disk* is a continuous map $\mathbb{B}^2 \to M$ sending $\mathbb{S}^1 = \partial \mathbb{B}^2$ to ∂M. Such a singular compression disk is *essential* if its restriction $\mathbb{S}^1 \to \partial M$ is not homotopic to 0 in ∂M. The celebrated Dehn's Lemma and Loop Theorem of Papakyriakopoulos [109] (see also Hempel [51]) assert that the existence of an essential singular compression disk is equivalent to that of an essential compression disk.

Note that a compression disk for ∂M gives a sphere in the double DM. An easy homology calculation shows that this sphere is essential if the compression disk is essential. With Theorem 2.2, this gives:

THEOREM 2.12. *If the* 3*-manifold M admits a complete geometric structure with totally geodesic boundary modelled over* $\mathbb{E}^3$, $\mathbb{S}^3$, $\mathbb{H}^3$, $\mathbb{H}^2 \times \mathbb{E}^1$ *or* Sol, *then M contains no essential* 2*-sphere, projective plane or compression disk.*

When M admits a Seifert-type geometry with totally geodesic boundary, the fibrations of the double DM provided by Theorem 2.5 must be invariant under the double involution τ. When DM inherits a Seifert fibration from the geometric structure, this Seifert fibration can project to either a Seifert fibration of M or to a locally trivial bundle structure on M whose fibers are compact intervals.

In the first case, a Seifert fibration of a manifold with boundary is defined as in the empty boundary case: The boundary is a union of fibers, and every boundary fiber admits a neighborhood U diffeomorphic to the quotient of $\mathbb{S}^1 \times \mathbb{B}^2_+$ by a free action of a finite group respecting the product structure, in such a way that the fibers of the fibration correspond to the image of the circles $\mathbb{S}^1 \times \{z\}$ and that $U \cap \partial M$ is the image of $\mathbb{S}^1 \times (\mathbb{B}^2_+ \cap i\mathbb{R})$. Here, $\mathbb{B}^2_+$ is the half-disk $\{z \in \mathbb{C};\ |z| \leqslant 1,\ \mathrm{Re}\, z \geqslant 0\}$ and $i\mathbb{R}$ is the imaginary axis. By inspection, such a neighborhood U of a boundary point must be diffeomorphic, either to $\mathbb{S}^1 \times \mathbb{B}^2_+$ with the product fibration, or to $[0,1] \times \mathbb{B}^2_+/\sim$ where $\sim$ identifies $\{0\} \times \mathbb{B}^2_+$ to $\{1\} \times \mathbb{B}^2_+$ by complex conjugation and where the fibers correspond to the sets $[0,1] \times \{z_0, \bar{z}_0\}$.

The space of leaves of a Seifert fibration is a 2-dimensional orbifold with boundary, where an orbifold with boundary is defined along the lines of the definition in the empty boundary case, allowing folding charts to originate from manifolds with boundary. The topological space underlying this base 2-orbifold is a surface with boundary, but its boundary points fall into two categories: Those which really are in the boundary of the orbifold, and correspond to fibers contained in the boundary of the 3-manifold; and those which really are in the interior of the orbifold, with isotropy group $\mathbb{Z}_2$ acting by reflection, and correspond to orientation reversing fibers in the interior of the 3-manifold.

The classification of Seifert fibered 3-manifolds with boundary is essentially the same as the one discussed in Section 2.4, and in particular Theorem 2.4. The Euler number $e_0 \in \mathbb{Q}/\mathbb{Z}$ is undefined when the boundary is non-empty.

THEOREM 2.13. *Let the* 3*-manifold M with non-empty boundary ∂M admit a complete geometric structure with totally geodesic boundary, modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$ or $\mathbb{E}^3 \cong \mathbb{E}^2 \times \mathbb{E}^1$. Then, possibly after a change of the splitting $\mathbb{E}^3 \cong \mathbb{E}^2 \times \mathbb{E}^1$, one of the following occurs*:

(i) *The foliation of the model space by the $\mathbb{E}^1$ factors projects to a Seifert fibration on M. In this case, the metric of the $\mathbb{S}^2$, $\mathbb{H}^2$ or $\mathbb{E}^2$ factors projects to a complete spherical, hyperbolic or euclidean, respectively, geometric structure with geodesic boundary on the base orbifold of the fibration.*

(ii) *The foliation of the model space by the $\mathbb{E}^1$ factors projects to a locally trivial bundle structure on M, with base a surface with boundary and with fiber an interval. In this case, the metric of the $\mathbb{S}^2$, $\mathbb{H}^2$ or $\mathbb{E}^2$ factors projects to a complete spherical, hyperbolic or euclidean, respectively, geometric structure with geodesic boundary on the base surface.*

(iii) *At most one leaf of the foliation of M by the $\mathbb{E}^1$ factors is a closed subset of M. In this case, M is diffeomorphic to one of the two $\mathbb{B}^2$-bundles over $\mathbb{S}^1$, or to an interval bundle over the* 2*-torus or the Klein bottle.*

In (ii), the interval fiber of the bundle can be open, compact or semi-open. It is compact precisely when the fibration of the double DM is a Seifert fibration, and the double involution τ acts by reflection on each of its fibers. In (iii), the interval fibers can be compact or semi-open.

In the 3-manifold M with boundary ∂M, let a *singular annulus* be a continuous map $\mathbb{A}^2 \to M$ sending the boundary $\partial\mathbb{A}^2$ to ∂M, where $\mathbb{A}^2$ denotes the standard annulus (or cylinder) $\mathbb{A}^2 = \mathbb{S}^1 \times [0,1]$. Let an *end* of M be a subset $U \subset M$ that is properly diffeomorphic to a manifold with corners $S \times [0,\infty[$, where S is a compact surface with boundary and where $U \cap \partial M$ corresponds to $(\partial S) \times [0, \infty[$. A singular annulus $\mathbb{A}^2 \to M$ is *essential* if the induced homomorphisms $\pi_1(\mathbb{A}^2) \to \pi_1(M)$ and $\pi_1(\mathbb{A}^2, \partial\mathbb{A}^2) \to \pi_1(M, \partial M)$ are injective and if it cannot be homotoped into an end $U \cong S \times [0, \infty[$ by a homotopy keeping the image of $\partial\mathbb{A}^2$ in ∂M; note that the property involving fundamental groups is independent of the choice of base point.

As in the case of singular tori, easy homotopic considerations show that, if the induced homomorphisms $\pi_1(\mathbb{A}^2) \to \pi_1(M)$ and $\pi_1(\mathbb{A}^2, \partial\mathbb{A}^2) \to \pi_1(M, \partial M)$ are injective and if the singular annulus $\mathbb{A}^2 \to M$ can be homotoped into an end $U \cong S \times [0, \infty[$ by a homotopy keeping the image of $\partial\mathbb{A}^2$ in ∂M, then the surface S must be an annulus or a Möbius strip. In particular, only *annular* ends $U \cong S \times [0, \infty[$, namely those for which S is an annulus or a Möbius strip, are relevant here.

Doubling the annulus $\mathbb{A}^2$ along its boundary gives a 2-torus. Therefore, a singular annulus $\mathbb{A}^2 \to M$ defines a singular 2-torus $\mathbb{T}^2 \to DM$ in the double DM of M. An easy cut and paste argument shows that this singular 2-torus is essential in DM if and only if the singular annulus is essential in M.

Applying Theorem 2.9 to the double DM, one easily obtains:

THEOREM 2.14. *If the 3-manifold M with boundary admits a complete hyperbolic structure with totally geodesic boundary, then M contains no essential 2-sphere, projective plane, compression disk, singular 2-torus or singular annulus. If, in addition, the hyperbolic structure of M has finite volume, then M is the union of a compact subset and of finitely many annular and toric ends.*

Similarly, an application of Theorem 2.11 gives:

THEOREM 2.15. *If the 3-manifold M with non-empty boundary admits a complete geometric structure modelled over* Sol, *then M is an interval bundle over the plane, the 2-torus, the Klein bottle, a compact, semi-open or open annulus, or a compact or open Möbius strip.*

In Theorem 2.15, the interval fiber can be compact, open or semi-open.

3. Characteristic splittings

We saw that a 3-manifold M very seldom admits a geometric structure modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \tilde{\times} \mathbb{E}^1$, $\mathbb{E}^2 \tilde{\times} \mathbb{E}^1$, $\mathbb{E}^3$, $\mathbb{S}^3$ or Sol. On the other hand, the only obstructions to the existence of a hyperbolic structure which we encountered consisted of essential 2-spheres, projective planes, singular 2-tori and, in the case with boundary, compression disks and singular annuli. In this section, we will see that a general 3-manifold M splits into pieces where these topological obstructions vanish. Most of these splittings are unique up to isotopy.

Recall that two diffeomorphisms $f_0, f_1 : M \to N$ are *isotopic* if they can be connected by a family of diffeomorphisms $f_t : M \to N$ which depend differentiably on $t \in [0, 1]$. Traditionally, the term *isotopy* refers to either a family of diffeomorphisms $f_t : M \to N$,

$t \in [0, 1]$, as above, or a diffeomorphism $f : M \to M$ which is isotopic to the identity. We will mostly use the second convention in this chapter.

3.1. *Connected sum decompositions*

First, let us focus on essential 2-spheres. Recall that a 2-sphere S embedded in the 3-manifold M is *essential* if it does not bound a 3-ball in M. The analysis of such essential spheres is equivalent to the Kneser–Milnor theory of connected sums of 3-manifolds. In addition to the original articles by Kneser [71] and Milnor [95], the book by Hempel [51] is an excellent reference for this material.

Let the 3-manifold M contain an essential 2-sphere S. If S is non-separating, there is a simple arc k embedded in $M - S$ which goes from one side of S to the other one. (Note that S always has two distinct sides, namely its normal bundle is trivial, because S is simply connected.) Thickening $S \cup k$, we obtain a 3-dimensional submanifold U of M bounded by an embedded 2-sphere S'. If S' is not essential, the closure of the complement of U in M is a 3-ball, and it is not too hard to see that M is diffeomorphic to one of the two $\mathbb{S}^2$-bundles over the circle $\mathbb{S}^1$; the topological type of the bundle so obtained depends on whether the arc k is orientation-preserving or -reversing. Therefore, if M contains an essential sphere, either it contains a separating essential sphere or it is diffeomorphic to one of the two $\mathbb{S}^2$-bundles over $\mathbb{S}^1$. This enables us to focus our attention on separating essential 2-spheres.

Now, let us introduce the additional hypothesis that the 3-manifold M is *of finite type*, namely is diffeomorphic to the interior of a compact manifold $\overline{M}$ with (possibly empty) boundary.

THEOREM 3.1 (Unique decomposition along 2-spheres). *Let M be a 3-manifold of finite type. Then, there is a compact 2-dimensional submanifold Σ of M such that:*

(i) *The components of Σ are separating 2-spheres in M.*

(ii) *If Σ is non-empty and if $M_0, M_1, \ldots, M_n$ are the closures of the components of $M - \Sigma$, then M_0 is diffeomorphic to a 3-sphere minus n finitely many disjoint open 3-balls; for every $i \geqslant 1$, M_i contains a unique component of Σ and every separating essential sphere in M_i is parallel to this component of Σ.*

(iii) *If M is non-orientable, no M_i is diffeomorphic to $\mathbb{S}^1 \times \mathbb{S}^2$ minus an open 3-ball.*

The family Σ is empty when M is the 3-sphere $\mathbb{S}^3$, and consists of a single sphere (bounding the 3-ball M_0) when M contains no separating essential 2-sphere. If Σ has at least two components, each of its components is an essential 2-sphere.

In addition, the list of the M_i is unique up to diffeomorphism. Namely, if Σ' is another family satisfying the same conditions and we use primes to denote the data associated to Σ', there are as many M_i' as M_i, and we can index these pieces so that each M_i' is diffeomorphic to the corresponding M_i. A stronger uniqueness property holds when M is orientable: If M is orientable, there exists an orientation-preserving diffeomorphism $\varphi : M \to M$ such that $\varphi(\Sigma) = \Sigma'$ and such that φ coincides with the identity outside of a certain compact subset of M.

In (ii), two surfaces Σ_1 and Σ_2 in M are *parallel* if they are separated by a component of $M - \Sigma_1 \cup \Sigma_2$ whose closure is diffeomorphic to the product $\Sigma_1 \times [0, 1]$.

A proof of the stronger uniqueness property of Theorem 3.1 for the orientable case, which the author learned from M. Scharlemann, can be found in Bonahon [16, Appendix A]. The rest of the statement is proved in Hempel [51, Chapter 3]. The existence of Σ was originally proved by Kneser [71], and the key ingredient is that there is a number n_0, depending only on the rank of $H_2(M; \mathbb{Z}/2)$ and on the number of simplices in a triangulation of $\overline{M}$, which bounds the number of components of a family Σ_1 of disjoint essential 2-spheres in M which are pairwise non-parallel. The uniqueness of the M_i was proved by Milnor [95].

The decomposition of a 3-manifold into simpler pieces which is provided by Theorem 3.1 has an inverse construction, defined by the operation of connected sum. More precisely, in the situation of Theorem 3.1, consider each piece M_i with $i \geqslant 1$, and let $\widehat{M}_i$ be obtained from M_i by gluing a 3-ball along the 2-sphere $\Sigma \cap M_i$; since every diffeomorphism of the 2-sphere extends to the 3-ball (see for instance [132,46]), this $\widehat{M}_i$ does not depend on the gluing map, up to diffeomorphism. Note that M_i is not a 3-ball since the 2-sphere ∂M_i is essential in M, and that $\widehat{M}_i$ consequently cannot be diffeomorphic to the 3-sphere (by Alexander's theorem [4]). Also, by the hypothesis that every essential separating 2-sphere in M_i is parallel to $\Sigma \cap M_i$, the manifold $\widehat{M}_i$ contains no separating essential 2-sphere. Therefore, Theorem 3.1 associates to any 3-manifold a finite collection of 3-manifolds $\widehat{M}_i$, $i = 1, \dots, n$, which contain no essential separating 2-sphere and which are not diffeomorphic to the 3-sphere.

Conversely, if we are given a finite collection of 3-manifolds $\widehat{M}_1, \dots, \widehat{M}_n$ of finite type, we can reconstruct a manifold M as follows. First remove the interior of a closed 3-ball B_i from each $\widehat{M}_i$, to obtain a manifold M_i bounded by a 2-sphere $\partial M_i = \partial B_i$; in addition, choose an orientation for each ball B_i. Then, remove from the 3-sphere the interiors of n disjoint closed balls $B'_1, \dots, B'_n$ to obtain a manifold M_0 bounded by n 2-spheres. And, finally, glue each M_i to M_0 by identifying each $\partial B_i \subset \partial M_i$ to $\partial B'_i \subset \partial M_0$ via an orientation-reversing diffeomorphism. Since any two orientation-preserving diffeomorphisms of the 2-sphere are isotopic (see for instance [132,46]), the resulting manifold M depends only on the manifold $\widehat{M}_i$ and on the oriented 3-balls B_i and B'_i. If at least one of the $\widehat{M}_i$ is non-orientable, sliding the balls around easily shows that M actually depends only on the manifolds $\widehat{M}_i$, up to diffeomorphism; by definition, M is the *connected sum* of the manifolds $\widehat{M}_1, \dots, \widehat{M}_n$. If all of the $\widehat{M}_i$ are orientable, some more care is required, and we need to choose an orientation for each of the $\widehat{M}_i$; then, if we insist that the balls B_i and B'_i are oriented by restriction of the orientations of $\widehat{M}_i$ and of the 3-sphere $\mathbb{S}^3$, the oriented manifold M does not depend on any other choice up to orientation-preserving diffeomorphism; by definition, the oriented manifold M is the *connected sum* of the oriented manifolds $\widehat{M}_i$. Reversing the orientation of one of the $\widehat{M}_i$ may change the diffeomorphism type of the resulting connected sum; see Hempel [51, Example 3.22].

The above operation of connected sum, combined with Theorem 3.1, provides a unique factorization of each 3-manifold of finite type as the connected sum of finitely many (oriented, if applicable) 3-manifolds $\widehat{M}_i$ which contain no essential separating 2-sphere and which are not diffeomorphic to the 3-sphere.

The classification problem is of course not the only problem topologists can be interested in. For instance, one is often led to analyze the topology of the group of all diffeomorphisms of a given manifold. We refer to Laudenbach [73] or Hendricks and Laudenbach [53] for an analysis of the diffeomorphism group of a 3-dimensional connected sum in terms of the diffeomorphism groups of its prime factors.

Having analyzed essential 2-spheres, we now turn our attention to essential projective planes. By the previous step, we can restrict the analysis to 3-manifolds that contain no essential 2-spheres. It turns out that we then have a much stronger uniqueness property for essential projective planes than for essential 2-spheres.

THEOREM 3.2 (Characteristic family of 2-sided projective planes). *Let M be a 3-manifold of finite type that contains no essential 2-sphere. Then, there is a compact 2-dimensional submanifold Π of M such that:*

(i) *Every component of Π is a 2-sided projective plane.*

(ii) *No two components of Π are parallel.*

(iii) *Every 2-sided projective plane in $M - \Pi$ is parallel in M to a component of Π.*

In addition, such a Π is unique up to isotopy of M.

A proof of Theorem 3.2 can be found in Negami [100].

If we split M open along the submanifold Π, we obtain a 3-manifold M' bounded by finitely many projective planes such that, by Condition (iii), every 2-sided projective plane in M' is parallel to a boundary component. However, in contrast to what we were able to do after splitting a 3-manifold along essential 2-spheres in Theorem 3.1, there is no way we can plug the boundary projective planes of M' to obtain a 3-manifold with no essential projective plane. Indeed, Poincaré duality with coefficients in $\mathbb{Z}_2$ (for instance) shows that there is no compact 3-manifold whose boundary consists of a single projective plane.

There is a way to overcome this difficulty, but it involves leaving the category of 3-manifolds and enlarging the scope of the analysis to 3-dimensional orbifolds. Although the projective plane $\mathbb{RP}^2$ does not bound any compact 3-manifold, it does bound a relatively simple orbifold. Namely considering $\mathbb{RP}^2$ as the quotient of the 2-sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ by the map $x \mapsto -x$, it bounds the 3-orbifold quotient of the 3-ball $\mathbb{B}^3 \subset \mathbb{R}^3$ by the same map $x \mapsto -x$. This orbifold has underlying topological space the cone over $\mathbb{RP}^2$, and its singular set consists of a single point, corresponding to the origin. If we plug each boundary component of M' with a copy of this 3-orbifold, the orbifold $\widehat{M}'$ obtained in this way does contain 2-sided projective planes, but these projective planes are inessential in the framework of the connected sum factorization of 3-orbifolds discussed in Section 3.6.

Therefore, Theorems 3.1 and 3.2 essentially reduce the study of finite type 3-manifolds to that of 3-orbifolds that contain no essential 2-sphere or projective plane. At this point, the switch to orbifolds is necessary only when the 3-manifold considered is non-orientable. We will first restrict attention to manifolds containing no essential 2-spheres or projective planes, and return to orbifolds in Section 3.6.

We also refer to Kalliongis and McCullough [68,69] for an analysis of the connected components of the diffeomorphism group of a 3-manifold which contains essential projective planes, in terms of the groups similarly associated to the pieces of the decomposition of Theorem 3.2.

As an aside, we should mention that we could have discussed singular 2-spheres in 3-manifolds, in the same way as we considered essential 2-tori and annuli in Section 2.4. More precisely, a *singular* 2-*sphere* in the 3-manifold M is a continuous map $\mathbb{S}^2 \to M$, and it is *essential* precisely when it is not homotopic to a constant map. In particular, the existence of an essential singular 2-sphere is equivalent to the non-triviality of the second homotopy group $\pi_2(M)$. The reason for the omission is the celebrated Sphere Theorem of Papakyriakopoulos [109] (extended to the non-orientable case by Epstein [34]; see also [51, Section 4.12]), which states that the existence of an essential singular 2-sphere implies the existence of an essential embedded 2-sphere or projective plane.

THEOREM 3.3 (Sphere Theorem). *If the* 3-*manifold M contains an essential singular* 2-*sphere, then it contains an essential* (*embedded*) 2-*sphere or a* 2-*sided projective plane.*

Note an unfortunate pitfall in this terminology if the Poincaré Conjecture does not hold: If there indeed exists a 3-manifold P which is homotopy equivalent to but not diffeomorphic to $\mathbb{S}^3$ and if M is an arbitrary 3-manifold which is not diffeomorphic to $\mathbb{S}^3$ then, in the connected sum of M and P, the connected sum 2-sphere is essential as an embedded 2-sphere but inessential as a singular 2-sphere. However, this phenomenon occurs only in this situation, and most likely never occurs in view of the strong evidence in favor of the Poincaré Conjecture.

3.2. *The characteristic torus decomposition*

In view of the previous section, consider now a 3-manifold M which contains no essential 2-sphere or projective plane. In Theorem 2.9, we saw that M must contain no essential singular 2-torus to admit a hyperbolic structure.

THEOREM 3.4 (Characteristic torus decomposition). *Let M be a* 3-*manifold of finite type which contains no essential sphere or projective plane. Then, up to isotopy, there is a unique compact* 2-*dimensional submanifold T of M such that*:

(i) *Every component of T is* 2-*sided, and is an essential* 2-*torus or Klein bottle.*
(ii) *Every component of* $M - T$ *either contains no essential embedded* 2-*torus or Klein bottle, or else admits a Seifert fibration.*
(iii) *Property* (ii) *fails when any component of T is removed.*

Here, a 2-torus T embedded in the 3-manifold M is *essential* if the inclusion map $T \to M$ is an essential singular 2-torus in M, in the sense of Section 2.4. A Klein bottle K embedded in M is *essential* if the composition of the orientation covering $\mathbb{T}^2 \to K$ and of the inclusion map $K \to M$ is an essential singular 2-torus.

When M is orientable, all components of the submanifold T are 2-tori, and Theorem 3.4 is therefore known as the Characteristic Torus Decomposition Theorem.

Theorem 3.4 was first announced by Waldhausen [147], and a complete proof was published by Johannson [63], Jaco and Shalen [60]; see also [58]. Actually, these authors are only considering the orientable case, but a proof of the general case (with no significant difference) can be found in Bonahon and Siebenmann [21]. The existence of the submanifold T is an easy consequence of the finiteness argument of Kneser [71] which we already encountered in Theorem 3.1; what is really important in Theorem 3.4 is its uniqueness up to isotopy. From a historical point of view, we should also mention H. Schubert's unique decomposition of a knot into its satellites [123], where a precursor of the 2-torus decomposition for knot complements first appeared.

It turns out that the Seifert pieces of the above decomposition 'absorb' all essential singular 2-tori in M, in the following sense:

THEOREM 3.5 (Classification of essential singular tori). *Let M and T be as in Theorem 3.4. Then, every essential singular 2-torus $\varphi : \mathbb{T}^2 \to M$ can be homotoped so that one of the following holds:*

(i) *The image of φ is contained in a Seifert fibered component of $M - T$, and φ is vertical with respect to the Seifert fibration in the sense that, at each $x \in \mathbb{T}^2$, the differential of φ at x sends the tangent plane $T_x\mathbb{T}^2$ to a plane in $T_{\varphi(x)}M$ that is tangent to the fiber passing through $\varphi(x)$; in this case, it is possible to choose a diffeomorphism between $\mathbb{T}^2$ and $\mathbb{S}^1 \times \mathbb{S}^1$ such that the restriction of φ to each circle $\{x\} \times \mathbb{S}^1$ is a covering map onto a fiber of the Seifert fibration.*

(ii) *M admits a Seifert fibration, and φ is horizontal with respect to the Seifert fibration in the sense that, at each $x \in \mathbb{T}^2$, the differential of φ at x sends the tangent plane $T_x\mathbb{T}^2$ to a plane in $T_{\varphi(x)}M$ which is transverse to the fiber passing through $\varphi(x)$. In this case, the composition of $\varphi : \mathbb{T}^2 \to M$ with the projection $M \to \Sigma$ to the base orbifold Σ of the Seifert fibration is an orbifold covering map; in particular, M is compact and Σ has orbifold Euler characteristic $\chi_{\mathrm{orb}}(\Sigma) = 0$. In addition, the Euler number e_0 of the Seifert fibration is either undefined (when M is non-orientable) or 0.*

(iii) *The image of φ is contained in a component T_0 of T, and φ factors as the composition of a covering map $\mathbb{T}^2 \to T_0$ and of the inclusion map $T_0 \to M$.*

The proof of Theorem 3.5 is significantly more difficult than that of Theorem 3.4. In the case where M is a Haken 3-manifold (see Section 3.5 for a definition of Haken 3-manifolds), Theorem 3.5 was, again, first announced by Waldhausen [147], and a complete proof was published by Johannson [63], Jaco and Shalen [60] (see also Feustel [38] for a related result, and [58]). However, the case where M is non-Haken was settled only recently, following work of Scott [124,125], Mess [92], Gabai [40], Casson and Jungreis [27].

An immediate corollary is the so-called Torus Theorem:

COROLLARY 3.6 (Torus Theorem). *Let M be a 3-manifold of finite type that contains no essential 2-sphere or projective plane. Suppose that there exists an essential singular 2-torus $\varphi : \mathbb{T}^2 \to M$. Then, M contains an essential embedded 2-torus or Klein bottle, or else M admits a Seifert fibration (or both).*

3.3. *The characteristic compression body*

We now return to the case with boundary. Let M be a 3-manifold with boundary, and assume that it is *of finite type*. This now means that M is obtained from a compact manifold $\overline{M}$

with boundary by removing a 2-dimensional compact submanifold (with boundary) of ∂M. We use here the convention that a codimension 0 submanifold is any subset bounded by a codimension 1 submanifold.

The analysis of Section 3.1 carries over without modifications, and we can therefore restrict attention to the case where M contains no essential 2-sphere or projective plane.

In Section 2.5, we saw that one restriction to the existence of a hyperbolic structure with totally geodesic boundary on M is that its boundary ∂M should not admit any essential compression disk. There is a theory of connected sums along disks which is very analogous to the connected sum factorization of Theorem 3.1. However, the uniqueness part of this factorization is here much improved, because it leads to a uniqueness up to isotopy. Indeed, in close analogy with the characteristic torus decomposition of the previous section, there is a characteristic submanifold of M which absorbs all compression disks of M.

Let a *compression body* be a 3-manifold obtained from a product $V_0 = S \times [0, 1]$ by gluing 2-handles along $S \times \{1\}$, and capping off with a 3-handle some of the boundary 2-spheres which may have appeared in the process. Namely, we start from $S \times [0, 1]$, where S is a surface of finite type, not necessarily connected. Then, we glue n copies of the product $\mathbb{B}^1 \times \mathbb{B}^2$ (= 2-handles) along n disjoint embeddings of the annulus $\mathbb{B}^1 \times \partial\mathbb{B}^2$ in $S \times \{1\}$; there is a natural way to smooth the corners in this construction to obtain a differentiable 3-manifold V_1 with the same homeomorphism type. Finally, we can glue 3-balls (= 3-handles) along some boundary components of V_1 which are 2-spheres and are not contained in $S \times \{0\}$. By definition, a compression body is any 3-manifold V obtained in this way.[5]

The boundary of a compression body V can be split into two pieces: the *external boundary* $\partial_e V$, which corresponds to $S \times \{0\}$, and the *internal boundary* $\partial_i V = \partial V - \partial_e V$. It is relatively easy to see that, for a connected compression body V with non-empty internal boundary $\partial_i V \neq \emptyset$, the above description of V as a thickened surface with handles can be chosen so that it involves only 2-handles and no 3-handles.

Consider a 3-manifold M of finite type with boundary. In addition, we assume that M contains no essential 2-sphere or projective plane.

If M admits any essential compression disk for its boundary ∂M, the argument of Kneser [71] provides a compact 2-submanifold D of M, whose components are compression disks for ∂M, and such that any compression disk for ∂M that is disjoint from D must be parallel to a component of D. This submanifold D is in general far from being unique. However, if we thicken the union $\partial M \cup D$, we get a compression body $V_1 \subset M$ whose external boundary $\partial_e V_1$ is equal to ∂M and whose internal boundary $\partial_i V_1$ admits no compression disk. Some components of the internal boundary $\partial_i V_1$ may be 2-spheres, which necessarily bound components of the closure of $M - V_1$ which are 3-balls by hypothesis on M. Let V be the compression body union of V_1 and of all these 3-balls components of the closure of $M - V_1$.

[5] The author often gets credited with the introduction of the term "compression body". This expression was actually coined by Larry Siebenmann, as a replacement for the inelegant "product with handles" used in preliminary versions of [16]. In retrospect, the term "hollow handlebody", reminiscent of the "hohlbretzel" already used by Waldhausen in [146], would probably be more appropriate to deal with situations where a compression body does not occur as a classifying object for compression disks.

THEOREM 3.7 (Characteristic compression body). *Let M be a 3-manifold with boundary, which is of finite type and contains no essential 2-sphere or projective plane. Then, up to isotopy, M contains a unique compression body V such that the external boundary $\partial_e V$ is equal to ∂M, such that the closure $\overline{M - V}$ contains no essential compression disk for its boundary $\partial_i V$, and such that no component of $\overline{M - V}$ is a 3-ball. In addition, any singular compression disk $(\mathbb{B}^2, \partial\mathbb{B}^2) \to (M, \partial M)$ can be homotoped inside V by a homotopy keeping the image of $\partial\mathbb{B}^2$ in ∂M.*

Theorem 3.7 is proved in Bonahon [16] in the orientable case, and the proof automatically extends to the non-orientable case. Note that, when M contains no essential compression disk and is not a 3-ball, the characteristic compression body V is just a collar neighborhood of the boundary, diffeomorphic to $\partial M \times [0, 1]$.

In the situation of Theorem 3.7, let M_0 the the closure of the complement $M - V$ in M. Then, M_0 contains no essential 2-sphere, projective plane, or compression disk for its boundary.

Conversely, let M_0 be a possibly disconnected 3-manifold of finite type that contains no essential 2-sphere, projective plane, or compression disk for its boundary. If we are given n disjoint embeddings of the two disks $\mathbb{B}^2 \times \partial\mathbb{B}^1$ in ∂M_0, we can use these pairs of disks to glue n copies of the product $\mathbb{B}^2 \times \mathbb{B}^1$ ($=$ 1-handles) to M_0. We then obtain a new 3-manifold M. It is not very hard to check that M contains no essential 2-sphere or projective plane. Note that each 1-handle $\mathbb{B}^2 \times \mathbb{B}^1$ provides a compression disk $\mathbb{B}^2 \times \{0\}$ for ∂M. It immediately follows from the construction that, if V is the characteristic compression body associated to M by Theorem 3.7, the closure of $M - V$ is diffeomorphic to M_0. This provides an inverse construction to the splitting defined by Theorem 3.7.

3.4. *The characteristic torus/annulus decomposition*

The previous characteristic splittings enable us to analyze essential 2-spheres, projective planes, singular 2-tori and compression disks in 3-manifolds with boundary. It remains to consider essential singular annuli.

THEOREM 3.8 (Characteristic torus/annulus decomposition). *Let M be a 3-manifold of finite type with boundary, which contains no essential 2-sphere, projective plane, or compression disk for its boundary. Then, up to isotopy, there is a unique compact 2-dimensional submanifold F of M such that:*

(i) *Every component F_1 of F is 2-sided, and is an essential 2-torus, Klein bottle, annulus or Möbius strip with $F_1 \cap \partial M = \partial F_1$.*

(ii) *For every component W of $M - F$, either W contains no essential embedded 2-torus, Klein bottle, annulus or Möbius strip, or W admits a Seifert fibration for which $W \cap \partial M$ is a union of fibers, or else W admits the structure of a $\mathbb{B}^1$-bundle over a surface of finite type such that the corresponding $\partial\mathbb{B}^1$-bundle is equal to $W \cap \partial M$.*

(iii) *Property* (ii) *fails when any component of F is removed.*

In addition, note that the ends of a Seifert fibered component W of $M - F$ all are of toric type, and can be delimited by 2-tori and Klein bottles in W; let T_W be the union *of 2-tori and Klein bottles delimiting those ends of W whose closure contain at least one annulus or Möbius strip component of F. Let T be the union of all 2-torus and Klein bottle components of F, and of all T_W as W ranges over all Seifert fibered components of $M - F$. Then, T is the characteristic 2-submanifold of the Characteristic Torus Decomposition Theorem 3.4 of the interior* $\mathrm{int}(M) = M - \partial M$.

Again, an annulus A in M is *essential* if the inclusion map $A \to M$ is an essential singular annulus, in the sense of Section 2.5. A Möbius strip A in M is *essential* if the composition of the orientation covering $\mathbb{A}^2 \to A$ and of the inclusion map $A \to M$ is an essential singular annulus.

When M is orientable, all components of the submanifold F are tori and annuli, and the above theorem is therefore known as the Characteristic Torus/Annulus Decomposition Theorem.

As in the case of the Characteristic Torus Decomposition Theorem 3.4, Theorem 3.8 was first announced by Waldhausen [147], and a complete proof was published by Johannson [63], Jaco and Shalen [60] (and see Bonahon and Siebenmann [21] for the details of a proof in the non-orientable case, where Theorem 3.8 is interpreted as a generalization of Theorem 3.4 to certain 3-orbifolds).

The fibered parts of the characteristic torus/annulus decomposition of Theorem 3.8 absorb all essential singular annuli, as indicated by the following theorem, whose proof can be found in [63,60,58].

THEOREM 3.9 (Classification of essential singular annuli). *Let M and F by as in Theorem 3.8. Then, every essential singular annulus* $\varphi : (\mathbb{A}^2, \partial\mathbb{A}^2) \to (M, \partial M)$ *can be homotoped so that one of the following holds*:

(i) *The image of* φ *is contained in a Seifert fibered component of* $M - F$, *and is vertical with respect to the Seifert fibration in the sense that the restriction of* φ *to each circle* $\mathbb{S}^1 \times \{y\} \subset \mathbb{S}^1 \times \mathbb{B}^1 = \mathbb{A}^2$ *is a covering map onto a generic fiber of the Seifert fibration.*

(ii) *The image of* φ *is contained in a* $\mathbb{B}^1$*-bundle component of* $M - F$, *and is vertical with respect to this bundle in the sense that the restriction of* φ *to each arc* $\{x\} \times \mathbb{B}^1 \subset \mathbb{S}^1 \times \mathbb{B}^1 = \mathbb{A}^2$ *is a diffeomorphism onto a fiber of the* $\mathbb{B}^1$*-bundle.*

(iii) *The image of* φ *is contained in an annulus or Möbius strip component A of F, and* φ *factors as the composition of a covering map* $\mathbb{A}^2 \to A$ *and of the inclusion map* $A \to M$.

Again a corollary of Theorem 3.9 is the following Annulus Theorem:

COROLLARY 3.10 (Annulus Theorem). *Let M be a* 3*-manifold of finite type which contains no essential* 2*-sphere or projective plane. Suppose that there exists an essential singular annulus* $\varphi : (\mathbb{A}^2, \partial\mathbb{A}^2) \to (M, \partial M)$. *Then, M contains an essential embedded annulus or Möbius strip.*

Indeed, if M contains an essential singular annulus and if V is its characteristic compression body, an easy cut and paste argument shows that the closure of $M - V$ also contains an essential singular annulus, and enables us to reduce the analysis to the case where M contains no essential compression disk for its boundary. Then Theorem 3.9 shows that, either the characteristic torus/annulus 2-submanifold F of Theorem 3.8 has a component which is an annulus or a Möbius strip, or F is disjoint from ∂M and at least one component V of $M - F$ is a $\mathbb{B}^1$-bundle or a Seifert fibered manifold with $V \cap \partial M \neq \emptyset$. In the second case, a suitable closed curve in the base of the $\mathbb{B}^1$-bundle, or a suitable arc in the base orbifold of the Seifert fibration, provides an essential embedded annulus or Möbius strip.

3.5. *Homotopy equivalences between Haken 3-manifolds*

Originally, the characteristic torus/decomposition was not developed as an obstruction to the existence of geometric structures, but as a tool to analyze homotopy equivalences between 3-manifolds. This important topic deserves a little digression here. The survey article of Waldhausen [146] is a good reference for this material.

We begin by defining a technically important class of 3-manifolds, called Haken manifolds. Let M be a compact 3-manifold with boundary, and let F be a compact 2-submanifold of M, so that $F \cap \partial M = \partial M$. When F is a 2-sphere, a projective plane or a compression disk, we already defined what it means for F to be essential. For all other types of surfaces, we say that F is *essential* if it is 2-sided and if the homomorphisms $\pi_1(F) \to \pi_1(M)$ and $\pi_1(F, \partial F) \to \pi_1(M, \partial M)$ induced by the inclusion map are all injective, and this for all choices of base points. Note that this is consistent with the definition used for 2-tori, Klein bottles, annuli and Möbius strips in previous sections.

By definition, a *compact Haken 3-manifold* is a compact 3-manifold with boundary which contains no essential 2-sphere or projective plane but which contains at least one essential surface. It can be shown that the last condition is unnecessary if the boundary ∂M is non-empty: If the compact 3-manifold M has non-empty boundary, it necessarily contains an essential surface, unless M is a 3-ball; see for instance [51, Chapter 13]. The crucial technical property of compact Haken 3-manifolds is that they give rise to a finite sequence $M = M_0, M_1, \dots, M_n$ of manifolds with boundary such that each M_i is obtained by splitting M_{i-1} along an essential surface and such that the last manifold is a disjoint union of 3-balls. Such a finite sequence lends itself well to inductive procedures, which makes it a very useful technical tool.

The starting point is the following result of Waldhausen [145], extended to non-orientable 3-manifolds by Heil [49,50] (see also Hempel [51]).

THEOREM 3.11. *Let $f: M \to N$ be a homotopy equivalence between compact 3-manifolds with boundary such that f restricts to a homeomorphism $f_{|\partial M}: \partial M \to \partial N$. Assume that M is Haken and that N contains no essential 2-sphere or projective plane. Then, f is homotopic to a diffeomorphism, by a homotopy fixing the restriction of f to the boundary.*

When M admits no essential compression disk for its boundary and N is not a trivial interval bundle over a surface, the requirement that f restricts to a homeomorphism between boundaries can be replaced by the weaker hypothesis that $f(\partial M) \subset \partial N$; the conclusion is then also weaker, and only provides a homotopy from f to a diffeomorphism (with no control on the boundary).

We should also mention the following related result, also proved by Waldhausen in [145].

THEOREM 3.12. *Let f_0, $f_1 : M \to M$ be two diffeomorphisms of a compact Haken 3-manifold which are homotopic. If M admits essential compression disks for its boundary, assume in addition that each stage of the homotopy sends ∂M to ∂M. Then, f_0 and f_1 are isotopic.*

Theorem 3.12 is the extension to Haken 3-manifolds of a celebrated result of Baer for surfaces (see for instance Epstein [34]). The combination of Theorems 3.11 and 3.12 says that, for two compact Haken 3-manifolds M and N, the space of diffeomorphisms $f : M \to N$ has the same number of connected components as the space of homotopy equivalences $f : M \to N$ (sending boundary to boundary in the presence of compression disks). This was later extended by Hatcher [48], who proved that these two spaces have the same homotopy type.

In Theorem 3.11, the requirement that $f(\partial M) \subset \partial N$ is crucial. Indeed, suppose that M contains an essential 2-sided annulus A with $A \cap \partial M = \partial A$. Split M along A and, in the split manifold, glue the two sides of A back together through the diffeomorphism $(x, t) \mapsto (x, 1-t)$ of $A \cong \mathbb{S}^1 \times [0, 1]$ to obtain a new 3-manifold N. The two manifolds M and N are both homotopy equivalent to the space obtained from M by collapsing each arc $\{x\} \times [0, 1]$ of $A \cong \mathbb{S}^1 \times [0, 1]$ to a point. However, M and N are in general not diffeomorphic; for instance, if the two boundary components of A lie in different boundary components of M, the two manifolds M and N will have a different number of boundary components. Note that this construction is analogous to the homotopy equivalence between the annulus and the Möbius strip.

This construction can be slightly generalized in the following way. Let A be a 2-sided 2-submanifold of M such that each component of A is an essential annulus or Möbius strip, and such that the closure V of some component of $M - A$ is homeomorphic to a $\mathbb{B}^2$-bundle over $\mathbb{S}^1$. Let N be obtained by replacing V in M by another $\mathbb{B}^2$-bundle W over $\mathbb{S}^1$, such that the components of A wrap around the $\mathbb{S}^1$ factor of the bundle the same number of times as in V. Then, M and N are both homotopy equivalent to the space obtained from M by collapsing each fiber of the $\mathbb{B}^2$-bundle structure of V to a point, but M and N are in general not diffeomorphic. Call such a homotopy equivalence $M \to N$ a *flip homotopy equivalence*. Johannson [63] proved that this is essentially the only counter-example:

THEOREM 3.13 (Homotopy equivalences between Haken 3-manifolds). *Let M and M' be two compact Haken 3-manifolds which admit no essential compression disk for their boundaries, and let $f : M \to M'$ be a homotopy equivalence. Consider the characteristic decomposition of M along a family F of 2-tori, Klein bottles, annuli and Möbius strips, as in Theorem 3.8, and let W be the union of a small tubular neighborhood of the annulus and Möbius strip components of F and of those components of $M - F$ which touch ∂M and admit the structure of a Seifert fibration or a $\mathbb{B}^1$-bundle. Let F' and W' be similarly defined in M'. Then, f can be homotoped so that:*

(i) $f^{-1}(W') = W$ *and* $f^{-1}(M' - W') = M - W$;
(ii) f *induces a homeomorphism from* $M - W$ *to* $M' - W'$;
(iii) f *induces a homotopy equivalence from* W *to* W'.
In addition, f *is homotopic to a product of flip homotopy equivalences* $M = M_0 \to M_1 \to \cdots \to M_n = M'$ *as above.*

An immediate corollary is the following.

COROLLARY 3.14. *If* $f : M \to N$ *is a homotopy equivalence between compact Haken 3-manifolds and if* M *contains no essential disk, annulus or Möbius strip, then* f *is homotopic to a homeomorphism.*

Finally, to completely analyze homotopy equivalences between compact Haken 3-manifolds, we need to understand what happens in the presence of essential compression disks.

First, note that the fundamental group $\pi_1(M)$ is the only non-trivial homotopy group of a Haken manifold M. Indeed, since M contains no essential 2-sphere or projective plane, the Sphere Theorem 3.3 shows that $\pi_2(M) = 0$. Since M contains an essential surface, its fundamental group is infinite, and its universal covering $\widetilde{M}$ is a non-compact 3-manifold with $\pi_1(\widetilde{M}) = 0$ and $\pi_2(\widetilde{M}) = \pi_2(M) = 0$. This implies that $\widetilde{M}$ is contractible by the Hurewicz Theorem, and that $\pi_n(M) = \pi_n(\widetilde{M}) = 0$ for every n.

In particular, the data of a homotopy equivalence $f : M \to M'$ between Haken 3-manifolds is equivalent to the data of the induced isomorphism $f_* : \pi_1(M) \to \pi_1(M')$.

If M is a compact Haken 3-manifold which contains essential compressions disks and if V is its characteristic compression body, $\pi_1(M)$ is isomorphic to the free product of the fundamental groups of the components of $\overline{M - V}$ and of some infinite cyclic groups. Note that each component of $\overline{M - V}$ is a compact Haken 3-manifold which admits no essential compression disk. Also, no component of $\overline{M - V}$ can have an infinite cyclic fundamental group by, for instance, Theorem 3.11.

As a consequence, it follows from Kurosh's theorem on the uniqueness of free product decompositions (see [81]) that a homotopy equivalence $f : M \to M'$ between compact Haken 3-manifolds, with respective characteristic compression disks V and V', induces a homotopy equivalence $g : \overline{M - V} \to \overline{M' - V'}$. The homotopy equivalence $g : \overline{M - V} \to \overline{M' - V'}$ is analyzed by Theorem 3.13. Then, understanding f is essentially a matter of comparing the way the handles of V and V' fit with respect to the components of $\overline{M - V}$ and $\overline{M' - V'}$.

A typical example is the following. Let S be a compact surface without boundary, not a 2-sphere or a projective plane, and let M be a 3-manifold with boundary obtained from the disjoint union of $S \times [0, 1]$ and of a 1-handle $\mathbb{B}^1 \times [0, 1]$ by identifying $\mathbb{B}^1 \times \{0, 1\}$ to two disjoint disks in $S \times \{0, 1\}$. Up to diffeomorphism, there are four manifolds M which are obtained in this way, according to whether M is orientable or not and to whether ∂M is connected or not. However, these four Haken 3-manifolds have the same homotopy type.

These homotopy equivalences between compact Haken 3-manifolds with essential compression disks or, more generally, homotopy equivalences of connected sums of such 3-manifolds, are analyzed in detail in Kalliongis and McCullough [66,67].

3.6. *Characteristic splittings of 3-orbifolds*

There is, for 3-dimensional orbifolds, a theory of characteristic splittings which closely parallels the one for 3-manifolds which we described in the preceding sections. There are several motivations for such an extension.

In Section 3.1, we already mentioned that 3-orbifolds constitute the natural framework for a theory of connected sums along projective planes. Namely, after splitting a non-orientable 3-manifold along characteristic 2-sided projective planes as in Theorem 3.2, we can plug the boundary projective planes so obtained with copies of the 3-orbifold $\mathbb{B}^3/\mathbb{Z}_2$, where $\mathbb{Z}_2$ acts by the antipodal map $x \mapsto x$ to obtain a 3-orbifold N. The methods of this section will then enable us to obtain for N characteristic splittings analogous to those of Sections 3.2, 3.3 and 3.4, to analyze essential 2-tori, compression disks and annuli in N and M.

Another reason is that we will see in Section 4.5 that there is an existence theorem for geometric structures on 3-orbifolds which is somewhat stronger than the current existence theorems for 3-manifolds. This existence theorem for 3-orbifolds requires the vanishing of certain topological obstructions, analogous to those of Section 2.4. As in the case of manifolds, when these topological obstructions do not vanish, they are best understood in terms of characteristic splittings similar to those of the preceding sections.

Also, we will see that the switch to orbifolds enables us to unify the cases with and without boundary, and in particular to consider the Torus/Annulus Decomposition of Section 3.4 as a special case of the simpler torus decomposition of Section 3.2.

However, the main benefit of these characteristic splittings of 3-orbifolds, and of the subsequent existence theorems for geometric structures, is that they provide more insight on properly discontinuous group actions on 3-manifolds.

The splitting theorems for 3-orbifolds are usually obtained by a simple word-by-word translation of the corresponding statements for manifolds. The guiding principle in establishing the dictionary is the following. The orbifold equivalent of an object of type T is such that, in an orbifold M/Γ which is the quotient of a 3-manifold M by a finite group Γ, an orbifold object of type T in the orbifold M/Γ is exactly the image of a Γ-invariant family of disjoint objects of type T in the manifold M. This principle will probably become clearer to the reader after we put it in practice.

Let a *sphere 2-orbifold* be the orbifold quotient of the unit 2-sphere $\mathbb{S}^2$ by a finite group of diffeomorphisms. By inspection or by using Proposition 2.6 on the geometrization of 2-orbifolds, this finite subgroup of diffeomorphisms is conjugate to a subgroup of the orthogonal group $O(3)$. It is an easy exercise to list all such sphere 2-orbifolds. For instance, exactly two of them have no singular points, namely the 2-sphere $\mathbb{S}^2$ and the projective plane $\mathbb{RP}^2$. The quotient of $\mathbb{S}^2$ by $\mathbb{Z}_p$ acting by rotations gives the orbifold whose underlying topological space is a 2-sphere, and whose singular set consist of two points each with isotropy group $\mathbb{Z}_p$ acting by rotations. Another example includes the quotient of $\mathbb{S}^2$ by the full symmetry group of the regular dodecahedron; its underlying topological space is a disk, its singular set is the boundary of this disk, and the non-trivial isotropy groups are all $\mathbb{Z}_2$ acting by reflection, except for exactly three singular points where the isotropy groups are the dihedral groups of respective orders 4, 6 and 10, acting in the standard way.

Similarly, a *ball 3-orbifold* is the orbifold quotient of the closed unit ball $\mathbb{B}^3$ by a finite group of diffeomorphisms. It follows from a deep theorem, the proof of the Smith conjecture [10] (see in particular [91]) that this finite group of diffeomorphisms is conjugate to a finite subgroup of the orthogonal group O(3). As a consequence every ball 3-orbifold is, in a suitably defined sense, a cone over its boundary, which is a sphere 2-orbifold.

Recall that an orbifold is *uniformizable* if it is isomorphic to the orbifold quotient of a manifold by a properly discontinuous group action. Otherwise, it is *non-uniformizable*. In Lemma 2.3, we saw that an orbifold which admits a complete geometric structure is necessarily uniformizable.

By an easy covering space argument (and compare Proposition 2.6), the only non-uniformizable 2-orbifolds Σ are those of the following list:

(i) the underlying space of Σ is a 2-sphere, and its singular set consists of a single point, where the isotropy group is $\mathbb{Z}_p$ acting by rotations with $p \geqslant 2$;

(ii) the underlying space of Σ is a 2-sphere, its singular set consists of two points, and their isotropy groups are $\mathbb{Z}_p$ and $\mathbb{Z}_q$ acting by rotations, with $p, q \geqslant 2$ distinct;

(iii) the underlying space of Σ is a 2-ball, the singular set is the boundary of this 2-ball, and all non-trivial isotropy groups are $\mathbb{Z}_2$ acting by reflection except at one point, where the isotropy group is the dihedral group of order $2p$ acting in the usual way, with $p \geqslant 2$;

(iv) the underlying space of Σ is a 2-ball, the singular set is the boundary of this 2-ball, and all non-trivial isotropy groups are $\mathbb{Z}_2$ acting by reflection except at two points, where the isotropy groups are the dihedral groups of order $2p$ and $2q$ acting in the usual way, with $p, q \geqslant 2$ distinct.

The existence of a geometric structure on the 3-orbifold M imposes restrictions on the 2-suborbifolds of M, where suborbifolds are defined in the obvious way: A *suborbifold* of the orbifold O is a subset S of the topological space underlying O such that, for every folding chart $f : \widetilde{U} \to U$ of the orbifold atlas O, $f^{-1}(\widetilde{U} \cap S)$ is a submanifold of U. Note that S inherits an orbifold structure by restriction of the charts of O.

By analogy with the case of 3-manifolds, a sphere 2-suborbifold S of M (namely a 2-dimensional suborbifold of M which is a sphere 2-orbifold) is *essential* if it is 2-sided and if it does not bound a ball 3-orbifold in M. Here, S is *2-sided* if it admits a neighborhood that is isomorphic to the orbifold $S \times \mathbb{B}^1$ or, equivalently, if its (suitably defined) orbifold normal bundle is trivial.

THEOREM 3.15. *If the 3-orbifold M admit a complete geometric structure, then every 2-suborbifold of M is uniformizable. If in addition, the geometric structure is modelled over any of the eight geometries of Section 2 except $\mathbb{S}^1 \times \mathbb{S}^2$, then M contains no essential sphere 2-suborbifold.*

The first statement comes from Lemma 2.3. Indeed, this result asserts that M is uniformizable, and any uniformization for M (namely an isomorphism between M and the orbifold X/Γ where the group Γ act properly discontinuously on the manifold X) provides a uniformization for any suborbifold of M.

The second statement is analogous to Theorem 2.2 and follows from the fact that the remaining model spaces X are irreducible. Indeed, Lemma 2.3 shows that M is isomorphic to the quotient orbifold X/Γ of the model space X by a properly discontinuous action of a group Γ of isometries. If S is a 2-sided sphere 2-suborbifold S of $M = X/\Gamma$, its pre-image in X is a Γ-invariant family of disjoint 2-spheres S_i in X. Since X is diffeomorphic to $\mathbb{R}^3$ or $\mathbb{S}^3$, an arbitrary component S_1 of this pre-image bounds a 3-ball B_1 in X. Then, if Γ_1 denotes the stabilizer of B_1 in Γ, the 2-suborbifold S bounds the ball 3-suborbifold B_1/Γ_1 in M. This shows that every 2-sided sphere 2-suborbifold of M bounds a ball 3-orbifold in M, and completes the proof of Theorem 3.15.

The requirement that the 3-orbifold M contains no essential sphere 2-suborbifold is stronger than might appear at first glance. When M is a manifold, namely when its singular set is empty, this condition holds if and only if it contains no essential 2-sphere and no 2-sided projective plane. Another fundamental case is when M is the mirror orbifold $DN/\mathbb{Z}_2$ associated to a 3-manifold N with boundary where, as in Section 2.5, DN is the double obtained by gluing two copies of N along their boundary and where $\mathbb{Z}_2$ acts on DN by exchange of these two copies. Then, the orbifold $M = DN/\mathbb{Z}_2$ satisfies this condition if and only if N admits no essential compression disks, 2-spheres, or projective planes.

There is no convenient characteristic splitting which would reduce the analysis of 3-orbifolds to those which contain no non-uniformizable 2-suborbifolds. Consequently, we have to introduce this hypothesis right away, and exclude from the analysis those 3-orbifolds which contain non-uniformizable 2-suborbifolds.

Let M be a 3-orbifold of finite type which contains no non-uniformizable 2-suborbifolds. Then M has a natural splitting as a connected sum of 3-orbifolds without essential sphere 2-suborbifolds, which closely parallels the splitting discussed in Section 3.1. Indeed, we can consider in M a finite family Σ of 2-sided sphere 2-suborbifolds which are pairwise disjoint, do not bound any ball 3-orbifold in M, and are pairwise not parallel in the sense that no two components of Σ are separated by a component of $M - \Sigma$ which is (orbifold) isomorphic to the product of a sphere 2-orbifold and of an interval. If the orbifold M is of finite type, namely is isomorphic to the interior of a compact 3-orbifold with boundary, the argument of Kneser again shows that there exists such a finite family Σ which is maximal for these properties. Then, cut M open along Σ, and glue a ball 3-orbifold B over each boundary component S of the 3-orbifold so obtained; namely B is a cone over S. By construction the (possibly disconnected) 3-orbifold $\widehat{M}$ so obtained contains no essential sphere 2-orbifold. The proof of Theorem 3.1 immediately generalizes to show:

THEOREM 3.16. *If M is a 3-orbifold of finite type which does not contain any non-uniformizable 2-suborbifold, the irreducible 3-orbifold $\widehat{M}$ associated to M by the above construction (and containing no essential sphere 2-suborbifold) is independent of the choice of Σ, up to orbifold isomorphism.*

(As indicated in Section 3.1, if we apply this splitting-gluing process to a 3-manifold M that contains 2-sided projective planes, the orbifold $\widehat{M}$ provided by Theorem 3.16 will have singular points, precisely two for each projective plane.)

Conversely, it is possible to reconstruct M from the orbifold $\widehat{M}$ through connected sum operations, although the situation is slightly more complicated than for manifolds. The connected sum of the 3-orbifolds M_1 and M_2 is defined as soon as we are given ball 3-suborbifolds $B_1 \subset M_1$, $B_2 \subset M_2$ and an isomorphism between B_1 and B_2. Then the connected sum $M_1 \# M_2$ is defined by gluing the orbifolds $M_i' = M_i - \mathrm{int}(B_i)$ along their boundaries, using the restriction of the isomorphism $B_1 \cong B_2$. In the case of manifolds, we only had to worry about orientations. The situation is somewhat more complex for orbifolds, because a finite type 3-orbifold M can contain several ball 3-suborbifolds which are isomorphic, but not equivalent by an ambient orbifold isomorphism of M, and because there usually are more isotopy classes of isomorphisms between ball 3-orbifolds than between ball 3-manifolds. However, there are only finitely many such ambient isomorphism types of ball 3-orbifolds in M, and finitely many isotopy classes of isomorphisms between them. Therefore, there are only finitely many possible connected sums $M_1 \# M_2$ of the finite type 3-orbifolds M_1 and M_2, and the combinatorics of these finitely many possibilities are easy to analyze. We also have to consider self-connected sums where $M_1 = M_2$, for instance to deal with a manifold M that contains a non-separating 2-sided projective plane, but this presents no significant difficulty. In particular, in Theorem 3.16, the 3-orbifold M can be recovered from $\widehat{M}$ up to a finite number of ambiguities, which are easily analyzed.

Because of this, we can now restrict attention to 3-orbifolds which contain no non-uniformizable 2-suborbifolds and no essential sphere 2-suborbifolds.

In Section 2.4, we saw that a geometric structure of Seifert type on a manifold often leads to Seifert fibration on the manifold. A similar phenomenon occurs for orbifolds. Actually, the situation is much simpler in the orbifold framework, because Seifert fibrations just correspond to locally trivial $\mathbb{S}^1$-bundles in the category of orbifolds, as we now explain.

Let F be a manifold (to simplify; a similar definition could be made where F is an orbifold). A (locally trivial) orbifold *F-bundle* consists of two orbifolds M and B and of a continuous map $p: M \to B$ between their underlying topological spaces such that, for every $x \in B$, there exists a neighborhood U of x in the topological space underlying B, a folding chart $f: \widetilde{U} \to U$ of B, and a folding chart $g: \widetilde{U} \times F \to p^{-1}(U)$ of M for which: the folding group G_g of g respects the product structure of $\widetilde{U} \times F$; the folding group G_f of f consists of those automorphisms of $\widetilde{U}$ which are induced by elements of G_g; the map p coincides with the map from $p^{-1}(U) = \widetilde{U} \times F / G_g$ to $U = \widetilde{U}/G_f$ that is induced by the projection $\widetilde{U} \times F \to \widetilde{U}$. Note that the folding groups G_f and G_g may be different since some elements of G_g may act by the identity on $\widetilde{U}$.

For such an orbifold F-bundle, the pre-images $p^{-1}(x)$ are the *fibers* of the bundle. Note that each fiber has a natural orbifold structure, and is orbifold covered by the manifold F. If x is a regular point of B, namely if the isotropy group of x is trivial, the fibers above nearby points are all orbifold isomorphic to $p^{-1}(x)$. If B is connected, the set of its regular points is connected, and we conclude that all fibers over regular points are isomorphic. By definition, the orbifold $p^{-1}(x)$ with x regular is the *generic fiber* of the bundle.

In particular, if we go back to the definition of a Seifert fibration, we see that a Seifert fibration on a 3-manifold M gives an orbifold $\mathbb{S}^1$-bundle over the base 2-orbifold B of the Seifert fibration. Conversely, by inspection of all possible local types, one easily sees that an orbifold $\mathbb{S}^1$-bundle $p: M \to B$ where M is a 3-manifold (with all isotropy groups trivial) defines a Seifert fibration on M.

Another important example occurs when we have a 3-manifold M with boundary and we consider the orbifold $DM/\mathbb{Z}_2$, where DM is the double obtained by gluing two copies of M along their boundary and where $\mathbb{Z}_2$ acts by exchange of the two copies. If this 3-orbifold $DM/\mathbb{Z}_2$ is endowed with the structure of an $\mathbb{S}^1$-bundle, then inspection now shows that there are two cases: Either the generic fiber is the manifold $\mathbb{S}^1$, and the bundle structure induces a Seifert fibration on M for which ∂M is a union of fibers and for which the base orbifold is a 2-orbifold B with boundary; in this case, the base orbifold of the $\mathbb{S}^1$-bundle is the quotient orbifold $DB/\mathbb{Z}_2$ where DB is the double obtained by gluing two copies of B along their boundary and $\mathbb{Z}_2$ acts by exchange of the two copies. Or the generic fiber is the 1-orbifold $\mathbb{S}^1/\mathbb{Z}_2$ where $\mathbb{Z}_2$ acts on $\mathbb{S}^1$ by reflection, the base orbifold B of the $\mathbb{S}^1$-bundle is a manifold, and the bundle structure induces a locally trivial (manifold) bundle on M with base the 2-manifold B and with fiber the interval $\mathbb{B}^1$, in such a way that ∂M corresponds to the $\partial\mathbb{B}^1$-subbundle.

The orbifold $\mathbb{S}^1$-bundles over a given 2-orbifold B are classified in Bonahon and Siebenmann [20]. The classification is very much in the spirit of Seifert's classification of Seifert fibrations [128]. It involves the consideration of the possible local types for such a bundle, plus certain global and semi-global invariants. The classification and its proof are not intrinsically more difficult than in the case of Seifert fibrations, but they are considerably more tedious because of the large number of possible local types.

As in Section 2.4, the existence of a complete Seifert-type geometry on a 3-orbifold M usually leads to a fibration on M. For simplicity, we restrict attention to finite volume structures.

THEOREM 3.17. *If the 3-orbifold M admits a complete geometric structure of finite volume modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \tilde{\times} \mathbb{E}^1$ or $\mathbb{E}^2 \tilde{\times} \mathbb{E}^1$, then at least one of the following occurs:*

(i) *The foliation of M by the $\mathbb{E}^1$ factors defines an orbifold $\mathbb{S}^1$-bundle structure on M with base orbifold a 2-orbifold B. In this case, the metric of the $\mathbb{S}^2$, $\mathbb{H}^2$ or $\mathbb{E}^2$ factors projects to a complete spherical, hyperbolic or euclidean structure of finite area on the base orbifold B.*

(ii) *The model space is $\mathbb{S}^2 \times \mathbb{E}^1$, and the $\mathbb{S}^2$-factors define an orbifold $\mathbb{S}^2$-bundle structure on M with base orbifold a compact 1-orbifold B.*

The proof is identical to that of the similar statement for manifolds, namely Theorem 2.5 in Section 2.4. The reader may want to back-track to that statement, and see how the list of compact exceptions in (iii) of Theorem 2.5 coincides with the list of all manifolds that correspond to the orbifold bundles of (ii) in the above Theorem 3.17.

The lucky coincidence which occurred for manifolds does not repeat for orbifolds: A geometric structure modelled on $\mathbb{E}^3$ or $\mathbb{S}^3$ for a 3-orbifold M does not necessarily produce a fibration. For instance, if Γ is the group consisting of all isometries of $\mathbb{E}^3$ that respect the lattice $\mathbb{Z}^3 \subset \mathbb{R}^3 = \mathbb{E}^3$, there is no identification of $\mathbb{E}^3$ with $\mathbb{E}^1 \times \mathbb{E}^2$ for which the $\mathbb{E}^1$ factors induce a fibration of the orbifold $\mathbb{E}^3/\Gamma$.

However, in the case of $\mathbb{E}^3$, the 3-dimensional crystallographic groups, namely the groups of isometries of $\mathbb{E}^3$ that act properly discontinuously and with compact quotient, were classified in the XIX century; see for instance Janssen [61] or Opechovski [103]. This classification is equivalent to the classification of all compact 3-orbifolds which admit a geometric structure modelled on $\mathbb{E}^3$. Up to orbifold isomorphism, there are 219 such compact euclidean 3-orbifolds (230 if we fix an orientation on orientable orbifolds). For noncompact euclidean 3-orbifold, one can rely on the celebrated theorem of Bieberbach [12,9] which asserts that, for every properly discontinuous group Γ of isometries of $\mathbb{E}^n$, there is an isometric splitting $\mathbb{E}^n \cong \mathbb{E}^p \times \mathbb{E}^q$ such that Γ respects some slice $\mathbb{E}^p \times \{x_0\}$ and $\mathbb{E}^p \times \{x_0\}/\Gamma$ is compact. Therefore, every non-compact euclidean 3-orbifold is an orbifold $\mathbb{E}^1$- of $\mathbb{E}^2$-bundle over a compact euclidean 2- or 1-orbifold. Since the 17 compact euclidean 2-orbifolds have been known for centuries (see for instance Montesinos [97]) and since there are only two compact euclidean 1-orbifolds ($\mathbb{S}^1$ and $\mathbb{S}^1/\mathbb{Z}_2$, where $\mathbb{Z}_2$ acts by reflection), this makes it an easy exercise to list all non-compact euclidean 3-orbifolds.

There is a similar classification for finite groups of isometries of $\mathbb{S}^3$; see Goursat [45], Threlfall and Seifert [137] and Du Val [32]. This is again equivalent to the classification of all spherical 3-orbifolds. For most of these spherical 3-orbifolds, the splitting $\mathbb{S}^3 = \mathbb{S}^2 \tilde{\times} \mathbb{S}^1$ induces an orbifold $\mathbb{S}^1$-bundle on this 3-orbifold, with basis a spherical 2-orbifold. However, several spherical 3-orbifolds admit no $\mathbb{S}^1$-bundle structure; see Dunbar [31].

THEOREM 3.18. *If the* 3*-orbifold M admits a complete geometric structure modelled over* $\mathbb{H}^3$, *it contains no essential* 2*-sided torus* 2*-suborbifold. In addition, it admits no structure as an orbifold* $\mathbb{S}^1$*-bundle, except in the case where the base* 2*-orbifold of the bundle is an open disk* 2*-orbifold or an open annulus* 2*-orbifold; this case cannot occur if the hyperbolic structure has finite volume.*

We have here used the automatic translation convention: A torus 2-suborbifold of M is a suborbifold which is isomorphic to the quotient of a 2-torus by a finite group action. Such a torus 2-suborbifold T is *incompressible* if, for every disk 2-suborbifold D (namely isomorphic to the quotient of a disk by a finite group action) in M with ∂D contained in T, there is a disk 2-suborbifold D' of T such that $\partial D' = \partial D$. A torus 2-suborbifold T of M is *essential* is it is incompressible and if it does not bound any end of M, namely if the closure of no component of $M - T$ is isomorphic to the orbifold $T \times [0, \infty[$. Incidentally, when M is a manifold, these conditions may seem weaker than the ones we considered in Section 2.4. However, for a 2-torus embedded in the 3-manifold M, they are actually equivalent by the Loop Theorem [109], Waldhausen's classification of incompressible surfaces in interval bundles [145, Proposition 3.1], and another lemma of Waldhausen [145, Proposition 5.4] which says that two disjoint incompressible surfaces are homotopic if and only if they are separated by the product of a surface with the interval.

For the Sol geometry, the same proof as in the manifold case of Theorem 2.11 yields the following result. See Scott [126] or Dunbar [30].

THEOREM 3.19. *If the* 3*-orbifold M admits a finite volume complete geometric structure modelled over* Sol, *it admits an orbifold fibration over the manifold* $\mathbb{S}^1$ *with generic fiber the* 2*-torus* $\mathbb{T}^2$ *or the orbifold* $\mathbb{T}^2/\mathbb{Z}_2$ *where* $\mathbb{Z}_2$ *acts on* $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ *by reflection on both* $\mathbb{S}^1$*-factors.*

In Theorem 3.18, we saw that essential sphere and torus 2-orbifolds are topological obstructions to the existence of hyperbolic structures on an orbifold. In this context, a quasi-automatic translation of the Torus Decomposition Theorem 3.4, proved in Bonahon and Siebenmann [21], gives:

THEOREM 3.20 (Characteristic torus decomposition for orbifolds). *Let M be a 3-orbifold of finite type which contains no non-uniformizable 2-suborbifold and no essential sphere 2-suborbifold. Then, up to orbifold isomorphism, there is a unique compact 2-dimensional suborbifold T of M such that:*

(i) *Every component of T is 2-sided, and is an essential torus 2-suborbifold.*
(ii) *Every component of $M - T$ either contains no essential torus 2-suborbifold, or admits an $\mathbb{S}^1$-bundle structure.*
(iii) *Property* (ii) *fails when any component of T is removed.*

As indicated earlier, in the case of a manifold M with boundary, applying this theorem to the double orbifold $DM/\mathbb{Z}_2$ subsumes both the Torus Decomposition Theorem 3.4 and the Torus/Annulus Decomposition Theorem 3.8 of Sections 3.2 and 3.4.

The same doubling trick allows us to avoid the study of 3-orbifolds with boundary.

Many other properties of 3-manifolds can be generalized to 3-orbifolds. For instance, see [136] for a generalization of Theorem 3.11 (on homotopy equivalences between Haken 3-manifolds) to a certain class of 3-orbifolds.

4. Existence properties for geometric structures

4.1. *The Geometrization Conjecture*

The central conjecture is that, for a 3-manifold of finite type, the topological obstructions considered in Sections 2.4 and 2.5 are the only obstructions to the existence of a complete geometric structure. If we combine this with the characteristic splittings of Section 3, this gives:

CONJECTURE 4.1 (*Geometrization Conjecture for* 3*-manifolds*). Let M be a 3-manifold of finite type with boundary which contains no essential 2-sphere, projective plane or compression disk. Let F be the 2-submanifold provided by the Characteristic Torus/Annulus Decomposition Theorem 3.8. Then, every component of $M - F$ admits a complete geometric structure with totally geodesic boundary.

If, in addition, M consists of a compact part and of finitely many toric or annular ends, then the geometric structures of the components of $M - F$ can be chosen to have finite volume, except in the following cases:

(i) $F = \emptyset$, M is non-compact, and M is diffeomorphic to an (open, closed or semi-open) interval bundle over a plane, an open annulus, an open Möbius strip, a 2-torus or a Klein bottle;
(ii) F consists of a single 2-torus, and $M - F$ is diffeomorphic to $F \times \mathbb{E}^1$; in this case, M is compact and admits a geometry modelled over Sol.

We will see in later sections that this conjecture is now proved in many important cases.

However, we should probably mention that the situation is still very unclear for 3-manifolds whose topological type is not finite. Indeed, some new topological obstructions then occur. For instance, the Whitehead manifold, a contractible 3-manifold which is not homeomorphic to $\mathbb{E}^3$ (see for instance Rolfsen [116, Section 3.I]), cannot admit a complete geometric structure since it is simply connected but is not homeomorphic to any of the model spaces. Some more topological obstructions related to finite topological type are discussed in Section 6.3. For 3-manifolds with finitely generated fundamental groups, it seems reasonable to conjecture that they can admit a complete geometric structure only if they have finite topological type (the so-called Marden Conjecture), which reduces the question to Conjecture 4.1. However, there is no clear conjecture for the manifolds with infinitely generated fundamental groups. This is in contrast to the case of surfaces, where complex analysis always provided a complete geometric structure.

4.2. *Seifert manifolds and interval bundles*

For the fibered pieces of the torus/annulus decomposition, the conclusions of the Geometrization Conjecture 4.1 can be proved by a relatively explicit construction; see Scott [126] or Kojima [72]. The proof is fairly simple for a Seifert fibration where the Euler number e_0 is undefined or 0, and requires just a little more care when $e_0 \neq 0$. It is convenient to consider, in addition to the four Seifert type geometries of Section 2.2, the non-maximal geometries of $\mathbb{E}^2 \times \mathbb{E}^1$ (contained in the geometry of $\mathbb{E}^3$) and $\mathbb{S}^2 \tilde{\times} \mathbb{S}^1$ (contained in the geometry of $\mathbb{S}^3$).

THEOREM 4.1 (Geometrization of Seifert fibered 3-manifolds). *Let the 3-manifold M with boundary admit a Seifert fibration with base 2-orbifold Σ. Let the orbifold Σ be endowed with a complete geometric structure. Let $l > 0$ be equal to $|e_0|\,\mathrm{area}(\Sigma)$ if the Euler number e_0 is defined (modulo a choice of orientation) and non-zero, and let $l > 0$ be arbitrary otherwise. Then, M admits a complete geometric structure modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{S}^2 \tilde{\times} \mathbb{S}^1$, $\mathbb{E}^2 \times \mathbb{E}^1$, $\mathbb{E}^2 \tilde{\times} \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$ or $\mathbb{H}^2 \tilde{\times} \mathbb{E}^1$, with totally geodesic boundary, in such a way that the $\mathbb{E}^1$ or $\mathbb{S}^1$ factors correspond to the fibers of the Seifert fibration, such that all generic fibers have length l, and such that the other factors project to the original geometric structure on the base orbifold Σ. In addition, if M is compact and oriented with $e_0 \neq 0$, the geometry is necessarily twisted. If M is compact and oriented with $e_0 = 0$, or compact and non-orientable, then the geometry is necessarily untwisted. If M is non-compact, the geometry can arbitrarily be twisted or untwisted.*

Note that the hypotheses of Theorem 4.1 are also necessary by Theorem 2.5.

Topologically, one might think that some Seifert fibered 3-manifolds are missing, namely those where the base 2-orbifold Σ admits no complete geometric structure. However, by Proposition 2.6, Σ then has underlying topological space a 2-sphere with 1 or 2 singular points. In this case, M is a lens space, and consequently admits a geometric structure modelled over $\mathbb{S}^2 \tilde{\times} \mathbb{S}^1$ for a different Seifert fibration; see [126,128,104,97].

Therefore, for the torus/annulus decomposition F of a finite type manifold M, any Seifert fibered component of $M - F$ really admits a complete geometric structure.

The proof of the similar statement for interval bundles is much simpler.

THEOREM 4.2 (Geometrization of 3-dimensional interval bundle). *Let the 3-manifold M with boundary admit the structure of an interval bundle over a surface S. Let Σ be endowed with a complete geometric structure. Then, M admits a complete geometric structure modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{E}^2 \times \mathbb{E}^1$ or $\mathbb{H}^2 \times \mathbb{E}^1$, with totally geodesic boundary, in such a way that the $\mathbb{E}^1$ factors correspond to the fibers of the bundle, and such that the other factors project to the original geometric structure on the base orbifold Σ.*

4.3. *The Hyperbolization Theorem for Haken 3-manifolds*

The most important theorem of this chapter certainly is the following existence theorem for hyperbolic structures on 3-manifolds.

In Section 3.5, we encountered the notion of compact Haken manifold. More generally, a *Haken manifold* is a 3-manifold which is obtained from a compact Haken manifold $\overline{M}$ by removing a compact 2-submanifold from $\partial \overline{M}$. In particular, a Haken manifold is a 3-manifold of finite type with boundary.

THEOREM 4.3 (Hyperbolization Theorem). *Let M be a Haken manifold which contains no essential 2-sphere, projective plane, 2-torus, compression disk or annulus. Then, M admits a complete hyperbolic structure with totally geodesic boundary.*

If, in addition, every end of M is toric or annular, then the complete hyperbolic structure has finite volume, unless M is diffeomorphic to an (open, closed or semi-open) interval bundle over a plane, an open annulus, an open Möbius strip, a 2-torus or a Klein bottle.

In other words, for Haken 3-manifolds, the conditions of Theorems 2.9 and 2.14 are necessary and sufficient for the existence of a complete hyperbolic structure with totally geodesic boundary (and possibly with finite volume). Combined with Theorems 4.1 and 4.2, Theorem 4.3 provides a proof of the Geometrization Conjecture 4.1 for all Haken 3-manifolds.

This theorem was first announced by Thurston around 1977; see [139,140]. The proof is very complex and, for a while, was not available in written form, although partial expositions such as those by Thurston [138,141,142], Morgan [98] or Sullivan [135] have been very influential in the further development of the field. Some detailed expositions of the proof of Theorem 4.3 are now beginning to become available. Technically, given an incompressible surface S in the Haken manifold M, the proof splits into two very distinct cases, according to whether or not a finite cover of M admits a structure of bundle over the circle $\mathbb{S}^1$ for which the pre-image of S is a union of fibers. The case of bundles over the circle is developed in detail in the monograph by Otal [106]. In the paper [107], Otal also gives a detailed proof of the non-bundle case, using a simplification of Thurston's original argument developed by McMullen [76,77]. The monograph by Kapovich [70] provides an exposition of the non-bundle case, following Thurston's original approach.

These proofs of the hyperbolization theorem are based on the idea, going back to Haken and Waldhausen, of successively cutting the 3-manifold along incompressible surfaces until one reaches a polyhedral ball. The characterization by Andreev [6,7] of the topological type of acute angle polyhedra in $\mathbb{H}^3$ enables one to put a hyperbolic structure on this polyhedral ball. The core of the proof is a difficult gluing process which, when the polyhedral ball is glued back together to reconstruct the original 3-manifold, progressively reconstructs a hyperbolic structure on the 3-manifold. Altogether, this approach is very reminiscent of the second method we used in Section 1.1 to construct hyperbolic structures on surfaces of finite type. A more analytic approach to the geometrization conjecture of 3-manifolds, such as the one proposed in [5] and in spirit closer to the first method of geometrization of surfaces discussed in Section 1.1, would certainly be more attractive but does not seem to be within reach at this point.

4.4. *Hyperbolic Dehn filling*

Hyperbolic Dehn filling is a method of constructing many hyperbolic manifolds by deformation of the structure of a complete hyperbolic 3-manifold with cusps. In addition to the original lecture notes by Thurston [138], the book by Benedetti and Petronio [11] is a good reference for the material in this section.

We begin with some topological preliminaries. Let M be a 3-manifold which is the interior of a compact 3-manifold $\overline{M}$ whose boundary $\partial\overline{M}$ consists of finitely many 2-tori. Let M' be a 3-manifold without boundary obtained as follows: Glue copies of the solid torus $\mathbb{S}^1 \times \mathbb{B}^2$ along some of the components of $\partial\overline{M}$, and remove the other components of $\partial\overline{M}$. Such a manifold M' is said to be obtained from M by *Dehn filling*.

For instance, a celebrated result of Lickorish and Wallace [74,151] says that every compact orientable 3-manifold can be obtained by Dehn filling along the complement of a link (= 1-submanifold) in the 3-sphere $\mathbb{S}^3$. See also the chapter by Boyer [22] for a more extensive discussion of Dehn fillings.

There are many possible ways of gluing a copy of the solid torus $\mathbb{S}^1 \times \mathbb{B}^2$ along a 2-torus component T of $\partial\overline{M}$. However, one easily sees that, up to diffeomorphism inducing the identity on $\overline{M}$, the resulting manifold is completely determined by the isotopy class of the simple closed curve $\{*\} \times \partial\mathbb{B}^2$ in T. In addition, this isotopy class is completely determined by the class of $H_1(T;\mathbb{Z})/\pm 1$ defined by $\{*\} \times \partial\mathbb{B}^2$, where ± 1 acts by multiplication on the homology group $H_1(T;\mathbb{Z})$ and where the ambiguity comes from the fact that the curve $\{*\} \times \partial\mathbb{B}^2$ is not oriented. See for instance Rolfsen's book [116, Chapter 9].

To specify a Dehn filling, one considers the boundary components $T_1, \dots, T_n$ of $\overline{M}$. The Dehn filling is then determined by the data of the *Dehn filling invariants* associated to these boundary 2-tori as follows: If a solid torus $\mathbb{S}^1 \times \mathbb{B}^2$ is glued along the 2-torus T_i, the Dehn filling invariant is the element $\delta_i \in H_1(T_i;\mathbb{Z})/\pm 1$ represented by the curve $\{*\} \times \partial\mathbb{B}^2$; if no solid torus is glued along T_i, the corresponding Dehn filling invariant is $\delta_i = \infty$, by definition. The motivation for this convention will become clear when we discuss the Hyperbolic Dehn Filling Theorem 4.4 below. Note that every list of indivisible elements of $\{\infty\} \cup H_1(T_i;\mathbb{Z})/\pm 1$ can be the list of Dehn filling invariants of some Dehn filling (an element $\delta \in H_1(T_i;\mathbb{Z})/\pm 1$ is *indivisible* if it cannot be written as $p\delta'$ for some integer $p \geqslant 2$ and $\delta' \in H_1(T_i;\mathbb{Z})/\pm 1$, and ∞ is indivisible by convention).

After these topological preliminaries, consider an orientable 3-manifold M which admits a complete hyperbolic metric of finite volume. By Theorem 2.9, it is diffeomorphic to the interior of a compact 3-manifold $\overline{M}$ whose boundary consists of 2-tori. By Lemma 2.3, the hyperbolic 3-manifold M is isometric to a quotient $\mathbb{H}^3/\Gamma$, where the group Γ acts properly discontinuously and by fixed point free isometries on the hyperbolic 3-space $\mathbb{H}^3$. By the theory of covering spaces, the group Γ is isometric to the fundamental group $\pi_1(M)$, and we therefore have an injective homomorphism $\rho_0: \pi_1(M) \to \mathrm{Isom}^+(\mathbb{H}^3)$ whose image is discrete, where $\mathrm{Isom}^+(\mathbb{H}^3)$ is the group of orientation-preserving isometries of $\mathbb{H}^3$.

We can actually put this in a more general framework. Consider on M a hyperbolic structure (or, more generally, an (X, G)-structure) which is not necessarily complete. Lift this structure to the universal covering $\widetilde{M}$ of M. Then, by following paths in $\widetilde{M}$, a relatively easy argument shows that every isometry from a small open subset of $\widetilde{M}$ to an open subset of $\mathbb{H}^3$ uniquely extends to a locally isometric map $D: \widetilde{M} \to \mathbb{H}^3$; see Thurston [138] or Benedetti and Petronio [11], and compare Singer [131]. This map D is a global isometry if and only if the hyperbolic metric is complete. From the uniqueness of the extension, we see that there is a homomorphism $\rho: \pi_1(M) \to \mathrm{Isom}^+(\mathbb{H}^3)$ such that $D(\gamma x) = \rho(\gamma) D(x)$ for every $x \in \widetilde{M}$ and $\gamma \in \pi_1(M)$. The map D is a *developing map* for the hyperbolic structure considered, and the homomorphism ρ is its *holonomy*.

Thurston observed that, if we consider the holonomy $\rho_0: \pi_1(M) \to \mathrm{Isom}^+(\mathbb{H}^3)$ of a complete hyperbolic structure, any homomorphism $\rho: \pi_1(M) \to \mathrm{Isom}^+(\mathbb{H}^3)$ that is sufficiently close to ρ_0 is the holonomy of a usually incomplete hyperbolic structure on M, which is of a very specific type near the ends of M. When this geometric structure is incomplete, its completion (as a metric space) is usually not a manifold. However, for some representations ρ near ρ_0, the completion of the geometric structure on M is a manifold, and, topologically, is obtained from M by Dehn filling. This enabled Thurston to prove the following theorem.

THEOREM 4.4 (Hyperbolic Dehn Filling Theorem). *Let the orientable 3-manifold M admit a finite volume complete hyperbolic structure. By Theorem 2.9, we know that there exists a compact 3-manifold $\overline{M}$ such that M is diffeomorphic to the interior of $\overline{M}$ and such that the boundary of $\partial \overline{M}$ consists of finitely many 2-tori $T_1, \dots, T_n$. Then, there is a finite subset X_i of each $H_1(T_i; \mathbb{Z})/\pm 1$ such that the following holds: If the manifold N is obtained from M by Dehn filling in such a way that the Dehn filling invariant $\delta_i \in \{\infty\} \cup H_1(T_i; \mathbb{Z})/\pm 1$ associated to each 2-torus T_i is not in X_i, then N admits a finite volume complete hyperbolic structure.*

As indicated above, the proof of Theorem 4.4 can be found in [138] or [11].

The proof of Theorem 4.4 also provides additional geometric information on the hyperbolic structure of the 3-manifold N. Then, for every compact subset K of M and every $\varepsilon > 0$, the hyperbolic metric of N can be chosen so that it is ε-close to the metric of M over K, provided that the Dehn filling invariants $\delta_i \in \{\infty\} \cup H_1(T_i; \mathbb{Z})/\pm 1$ are all sufficiently close to ∞. In addition, when $\delta_i \neq \infty$, we can arrange that the core $\mathbb{S}^1 \times \{0\}$ of the solid torus $\mathbb{S}^1 \times \mathbb{B}^2$ glued along the 2-torus T_i is a closed geodesic of N whose length tends to 0 as δ_i tends to ∞ in $\{\infty\} \cup H_1(T_i; \mathbb{Z})/\pm 1$ (for the topology of the 1-point compactification, for which the neighborhoods of ∞ are the complements of finite subsets of $H_1(T_i; \mathbb{Z})/\pm 1$).

The requirement that M is orientable is not crucial in Theorem 4.4. When M is non-orientable and admits a finite volume complete hyperbolic metric, we know from Theorem 2.9 that M is the union of a compact part and of finitely many ends, each diffeomorphic to $\mathbb{T}^2 \times [0, \infty[$ or $\mathbb{K}^2 \times [0, \infty[$, where $\mathbb{T}^2$ and $\mathbb{K}^2$ respectively denote the 2-torus and the Klein bottle. Again, any homomorphism $\rho : \pi_1(M) \to \mathrm{Isom}(\mathbb{H}^3)$ near the holonomy of the complete hyperbolic structure of M is the holonomy of a possibly incomplete hyperbolic structure on M, and we can consider the completion of this metric. However, this completion can almost never be a hyperbolic manifold near the Klein bottle ends of M. Nevertheless, this method provides a complete finite volume metric on any 3-manifold N obtained by sufficiently complicated Dehn filling on toric ends of M, leaving the Klein bottle ends topologically untouched.

There is also a version of Theorem 4.4 where we allow the complete hyperbolic metric of M to have infinite volume, provided we require the metric of M to be "geometrically finite"; see Comar [28] or Bonahon and Otal [19]. Again, it enables one to construct a complete hyperbolic metric on any 3-manifold N obtained by sufficiently complicated Dehn filling on the toric ends of M.

One of the drawbacks of Theorem 4.4 is that it is not explicit. Namely, it does not provide a method to determine the exceptional sets X_i, and not even an estimate on their sizes. There is strong experimental evidence that these X_i should be relatively small.

In some cases, it is possible to carry out explicitly the procedure of hyperbolic Dehn filling. A celebrated example is that of the complement of the figure eight knot in $\mathbb{S}^3$, investigated by Thurston in [139]. In this example, Thurston was able to analyze a large portion of the space of homomorphisms $\rho : \pi_1(M) \to \mathrm{Isom}(\mathbb{H}^3)$, and to show that certain 'integer points' in this space defined hyperbolic 3-manifolds obtained Dehn filling on the figure eight knot complement. He also observed that, as one approached some points of the boundary of this domain, the hyperbolic structures degenerated to Seifert-type geometric structures. The analysis of this example was instrumental in the development of the Geometrization Conjecture 4.1. A similar procedure is implemented for many link complements, as well as punctured 2-torus bundles over the circle, in the software SnapPea discussed in Section 6.1, and gives explicit upper bounds for the exceptional sets X_i in the examples considered.

There is also a more theoretical evidence for the X_i to be small, obtained by leaving the world of geometric structures. Indeed, instead of trying to put a hyperbolic structure on the manifold N obtained from M by Dehn filling, we can just try to endow it with a complete metric of negative curvature. M. Gromov and W. Thurston provided a technique to achieve this for most Dehn fillings.

More precisely, let M be an orientable 3-manifold with a finite volume complete hyperbolic structure. Then, each end of M has a neighborhood U which is isometric to a model H/Γ_0 where, in the hyperbolic 3-space $\mathbb{H}^3 = \{(u, v, w) \in \mathbb{R}^3;\ w > 0\}$, H is a *horoball* of the form $H = \{(u, v, w) \in \mathbb{R}^3;\ w \geqslant w_0\}$ for some positive constant w_0, and where Γ_0 is a group of horizontal translations in $\mathbb{H}^3 \subset \mathbb{R}^3$ which is isomorphic to $\mathbb{Z}^2$; see for instance [138, Chapter 5] or [11, Section D.3]. The 2-torus ∂U then is isometric to the quotient under Γ_0 of the plane of equation $w = w_0$. The hyperbolic metric of $\mathbb{H}^3$ induces a Euclidean metric on this plane, and therefore on the 2-torus ∂U. Let $\overline{M}$ be the compact manifold with boundary obtained by removing from M all the neighborhoods $U_1, \dots, U_n$ so associated to the ends of M. Note that M is diffeomorphic to the interior of $\overline{M}$, so that we can use $\overline{M}$ to define Dehn fillings along the ends of M. Let T_i denote the component $T_i = \partial U_i$ of $\partial \overline{M}$.

Note that, in the Euclidean 2-torus T_i, every non-trivial homology class in $H_1(T_i; \mathbb{Z})$ can be realized by a closed geodesic and that all closed geodesics in the same homology class have the same length.

THEOREM 4.5 (2π-theorem). *For M, $\overline{M}$ and $\partial\overline{M} = \bigcup_{i=1}^{n} T_i$ as above, let N be obtained from M by Dehn filling with Dehn filling invariants $\delta_i \in \{\infty\} \cup H_1(T_i; \mathbb{Z})/\pm 1$. Suppose that, whenever the invariant δ_i is not ∞, the class $\delta_i \in H_1(T_i; \mathbb{Z})/\pm 1$ can be realized in the euclidean 2-torus T_i by a closed geodesic of length strictly greater than 2π. Then, the 3-manifold N obtained by Dehn filling M according to the Dehn filling invariants δ_i admits a complete Riemannian metric whose curvature is bounded between two negative constants.*

The proof of Gromov and Thurston [47] is based on an explicit construction which extends the hyperbolic metric on $\overline{M} \subset M$ to a negatively curved metric on N.

A packing argument shows that the neighborhoods U_i of the ends of M can be chosen so that no closed geodesic of $T_i = \partial U_i$ has length less than 1. An elementary argument then bounds the number of homology classes of $H_1(T_i; \mathbb{Z})$ whose closed geodesic representatives have length at most 2π. This argument eventually gives the following corollary of Theorem 4.5, whose proof can be found in Bleiler and Hodgson [13].

THEOREM 4.6. *Let M be any orientable 3-manifold with a complete hyperbolic structure of finite volume, diffeomorphic to the interior of a compact manifold $\overline{M}$, and let the 2-tori $T_1, \ldots, T_n$ be the boundary components of $\overline{M}$. Then, there are finite subsets $X_i \subset H_1(T_i; \mathbb{Z})/\pm 1$, each with at most 48 elements, such that any manifold obtained from M by Dehn filling whose Dehn filling invariants are not in the X_i admits a complete Riemannian metric of negative curvature. When the boundary of $\overline{M}$ is connected, the exceptional set X_1 can be chosen to have at most 24 elements.*

The proof of Theorem 4.6 gives some additional information on the shape of the exceptional sets X_i; see Bleiler and Hodgson [13].

The main interest of Theorems 4.5 and 4.6 is that, if a 3-manifold N admits a complete Riemannian metric whose curvature is bounded between two negative constants, then N must satisfy the same topological restrictions as those given in Theorem 2.9 for the existence of a complete hyperbolic structure. Therefore, if the Geometrization Conjecture is true, any 3-manifold obtained by Dehn surgery as in Theorems 4.5 and 4.6 will also admit a complete hyperbolic metric. This gives a conjectural estimate on the size of the exceptional sets X_i in the Hyperbolic Dehn Filling Theorem 4.4.

There is evidence that the exceptional sets should actually be smaller than predicted by Theorem 4.6. See the chapter by Boyer [22] for a summary of what is currently known in this direction.

4.5. *Geometrization of 3-orbifolds*

In 1982, Thurston announced a proof of the following result.

THEOREM 4.7 (Geometrization Theorem for 3-orbifolds). *Let M be a 3-orbifold of finite type, which contains no non-uniformizable 2-suborbifold and no essential sphere 2-sub-*

orbifold, and let T be the characteristic torus 2-suborbifold provided by Theorem 3.20. Assume in addition that the singular set of M is non-empty and has dimension at least 1. Then, every component of $M - T$ admits a complete geometric structure.

COROLLARY 4.8. *Let M be a 3-orbifold of finite type which contains no non-uniformizable 2-suborbifold, and no essential sphere or torus 2-suborbifold. Assume in addition that the singular set of M is non-empty and has dimension at least 1. Then, M admits a complete geometric structure, modelled over one of the eight 3-dimensional geometries of Section 2.*

Expositions of Theorem 4.7 have only begun to appear in recent months. A complete exposition can be found in Cooper, Hodgson and Kerckhoff [29], while Boileau and Porti [15] is restricted to the important case where the singular locus consists of disjoint circles. Earlier partial results can be found in [55,134,156]; see also [87] for the case where the orbifold admits a finite orbifold covering which is a Seifert fibered manifold.

Theorem 4.7 has the following important corollary.

COROLLARY 4.9 (Geometrization of 3-manifolds with symmetries). *Let M be a 3-manifold of finite type which contains no essential 2-sphere, projective plane or 2-torus. Suppose that there exists a periodic diffeomorphism $f: M \to M$ whose fixed point set has dimension at least 1. Then, M admits a complete geometric structure.*

Corollary 4.9 follows by application of Theorem 4.7 to the orbifold $M/\mathbb{Z}_n$, where the action of the cyclic group $\mathbb{Z}_n$ is generated by f.

5. Uniqueness properties for geometric structures

5.1. *Mostow's rigidity*

The most important feature of hyperbolic structures is their uniqueness properties, which follows from Mostow's Rigidity Theorem [99]. This result deals with lattices of $\mathbb{H}^n$ (or, more generally, of any rank 1 homogeneous space), namely discrete groups Γ of isometries of $\mathbb{H}^n$ such that the quotient $\mathbb{H}^n/\Gamma$ has finite volume. The Rigidity Theorem of Mostow asserts that, for every group isomorphism $\varphi: \Gamma_1 \to \Gamma_2$ between two such lattices Γ_1 and Γ_2, there is an isometry $F: \mathbb{H}^n \to \mathbb{H}^n$ such that $\varphi(\gamma_1) = F\gamma_1 F^{-1}$ for every $\gamma_1 \in \Gamma_1 \subset \mathrm{Isom}(\mathbb{H}^n)$. A proof of this deep result can be found in the monograph by Mostow [99]. Other proofs appear in Thurston [138] or Benedetti and Petronio [11].

We can apply this result to the case of a complete hyperbolic 3-manifold M with totally geodesic boundary and with finite volume. As in Section 2.5, the hyperbolic structure of M gives a hyperbolic structure on the double DM, obtained by gluing two copies of M along their boundary. In particular, DM is isometric to the quotient of $\mathbb{H}^3$ by the properly discontinuous action of a group Γ' of isometries. We can then consider the uniform lattice Γ generated by Γ' and by any lift of the isometric involution which exchanges the two copies of M in $DM \cong \mathbb{H}^3/\Gamma'$. By construction, M is isometric to the quotient $\mathbb{H}^3/\Gamma$.

When we have two such 3-manifolds $M_1 \cong \mathbb{H}^3/\Gamma_1$ and $M_2 \cong \mathbb{H}^3/\Gamma_2$, any diffeomorphism $\varphi: M_1 \to M_2$ lifts to a diffeomorphism $\Phi: \mathbb{H}^3 \to \mathbb{H}^3$ which conjugates the action of Γ_1 to the action of Γ_2. Mostow's Rigidity Theorem provides an isometry $F : \mathbb{H}^3 \to \mathbb{H}^3$ which also conjugates Γ_1 to Γ_2 and induces the same isomorphism $\Gamma_1 \to \Gamma_2$. In particular, F induces an isometry $f : M_1 \to M_2$. If we identify M_1 to one of the two halves of the double DM_1, the fact that F and Φ act similarly on the corresponding subgroup $\pi_1(M_1) \subset \Gamma_1$ shows that f is homotopic to φ. This proves:

THEOREM 5.1 (Hyperbolic Rigidity Theorem). *Let M_1 and M_2 be two complete hyperbolic 3-manifolds with totally geodesic boundary and with finite volume. Then, every diffeomorphism $\varphi : M_1 \to M_2$ is homotopic to an isometry f.*

We can complement this theorem by adding that f and φ are actually isotopic. This follows from Waldhausen's Isotopy Theorem 3.12 when M is Haken, and in particular when M is non-compact or has non-empty boundary, and from the recent work of D. Gabai and collaborators which will be discussed in Section 6.4 in the general case.

An important practical corollary of Theorem 5.1 is that, if the 3-manifold M admits a finite volume complete hyperbolic metric with totally geodesic boundary, any geometric invariant of this hyperbolic metric is actually a topological invariant of M. Simple examples of such geometric invariants include the volume, or the (locally finite) set of lengths of the closed geodesics of the metric. More elaborate examples involve the Chern–Simons invariant [93], its refinement the eta-invariant [155,94,108], or the Ford domain discussed in Section 6.1.

5.2. *Seifert geometries and* Sol

We now consider the Seifert geometries. Given a 3-manifold M, we would like to classify the complete geometric structures modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$ or $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$ with which M can be endowed, up to isotopy. Namely, such a geometric structure is identified to an atlas which locally models M over the model space, and is maximal among all atlases with this property. We identify two such geometrical structures when the corresponding maximal atlases differ only by composition with a diffeomorphism of M which is isotopic to the identity.

We first restrict attention to the cases where, as in Theorem 2.5, the $\mathbb{E}^1$ factors of these geometries induce a Seifert fibration or a locally trivial bundle with fiber $\mathbb{E}^1$. We then split the problem in two parts: Classify all such fibrations of M, up to isotopy, and then, for a given fibration, classify the geometric structures which give this fibration. We can also add the model spaces $\mathbb{E}^2 \times \mathbb{E}^1$ and $\mathbb{S}^2 \widetilde{\times} \mathbb{S}^1$ to the geometries considered since, by Theorems 2.7 and 2.8, such geometric structures usually arise from geometric structures modelled over $\mathbb{E}^3$ and $\mathbb{S}^3$.

When a 3-manifold admits a Seifert fibration, this fibration is usually unique.

THEOREM 5.2 (Topological uniqueness of Seifert fibrations). *Let the finite type 3-manifold M admit a Seifert fibration with base 2-orbifold Σ. Suppose that the orbifold Euler characteristic $\chi_{\mathrm{orb}}(\Sigma)$ of Σ is non-positive; when $\chi_{\mathrm{orb}}(\Sigma) = 0$, suppose in addition that the manifold M is compact and orientable, and that the Euler number $e_0 \in \mathbb{Q}$ is non-trivial. Then, the Seifert fibration of M is unique up to isotopy.*

Theorem 5.2 was proved by Waldhausen [144] for Haken manifolds, by Scott [127] for most non-Haken manifolds and by Boileau and Otal [14] for the remaining cases.

The hypotheses that $\chi_{\mathrm{orb}}(\Sigma) \leqslant 0$ and $e_0 \neq 0$ when $\chi_{\mathrm{orb}}(\Sigma) = 0$ in Theorem 5.2 are necessary because the corresponding 3-manifolds may admit several non-isotopic Seifert fibrations. For instance, when the base 2-orbifold Σ has underlying topological space the 2-sphere $\mathbb{S}^2$ with $\leqslant 2$ singular points (in which case $\chi_{\mathrm{orb}}(\Sigma) > 0$), the manifold M is a lens space and admits many Seifert fibrations of the same type. Similarly, when Σ is the 2-torus manifold $\mathbb{T}^2$ with no singular point (in which case $\chi_{\mathrm{orb}}(\Sigma) = 0$) and $e_0 = 0$, M is the 3-torus $\mathbb{T}^3 = \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$, which admits many non-isotopic fibrations as a locally trivial $\mathbb{S}^1$-bundle. A more exotic example occurs when Σ has underlying topological space $\mathbb{S}^2$ with 3 singular points, two of which have isotropy group $\mathbb{Z}_2$; in this case M admits another Seifert fibration, whose base 2-orbifold is the projective plane with 0 or 1 singular point.

However, in the cases where Theorem 5.2 does not apply, Orlik, Vogt and Zieschang [105] proved a classification of the Seifert fibrations of M up to diffeomorphism of M that is homotopic to the identity; see also Orlik [104]. The work of various authors [8,14,17, 56,57,118,120] (see Boileau and Otal [14] for a historical guide through these references) later proved that every diffeomorphism of M that is homotopic to the identity is actually isotopic to the identity when M is orientable; when M is non-orientable, then it contains a 2-sided essential 2-sphere, projective 2-plane, 2-torus or Klein bottle, and the machinery used in the proofs of Theorems 3.1, 3.2 and 3.4 easily proves the same result in this case. Therefore, the classification of [105] is actually a classification of Seifert fibrations up to isotopy. This classification has too many cases to be listed here, and we can only refer the reader to [105] and [104] for precise statements.

The situation is simpler for $\mathbb{E}^1$-fibrations. The only minor complication occurs when the base of the fibration is non-compact. The example to keep in mind here is that the trivial $\mathbb{E}^1$-bundles over the 2-torus minus 1 point and over the 2-sphere minus 3 points have diffeomorphic underlying spaces. To deal with this problem, it is convenient to consider the compact interval $\widehat{\mathbb{E}}^1$ obtained by adding an end point to each of the two ends of $\mathbb{E}^1$. Then, if M admits a (locally trivial) $\mathbb{E}^1$-fibration with base a surface S, this fibration canonically extends to an $\widehat{\mathbb{E}}^1$-fibration with the same base S, whose underlying space $\widehat{M}$ is the union of M and of a 2-fold covering of S.

If the 3-manifold M has finite type, it is the interior of a compact 3-manifold $\overline{M}$ with boundary. It is an easy consequence of Waldhausen's collar lemma [145, Lemma 3.5] that this compactification $\overline{M}$ is unique up to diffeomorphism whose restriction to M is isotopic to the identity. Since $\pi_1(S) = \pi_1(M)$ is finitely generated, S is also of finite type. Considering a compactification of S, we conclude from the uniqueness of $\overline{M}$ that we can isotop the $\mathbb{E}^1$-fibration so that the associated $\widehat{\mathbb{E}}^1$-fibration space $\widehat{M}$ is obtained from $\overline{M}$ by removing disjoint annuli and Möbius strips from the boundary $\partial\overline{M}$, one for each end of S. Note that $\widehat{M} = \overline{M}$ when S is compact.

Waldhausen [145, Lemma 3.5] proved the following uniqueness result.

THEOREM 5.3 (Topological uniqueness of $\mathbb{E}^1$-fibrations). *Let the 3-manifold M be the interior of a compact 3-manifold $\overline{M}$ with boundary, and and consider two $\mathbb{E}^1$-fibrations* *whose associated $\widehat{\mathbb{E}}^1$-fibration spaces $\widehat{M}$, $\widehat{M}' \subset \overline{M}$ are isotopic in $\overline{M}$. Then, these two $\mathbb{E}^1$-fibrations of M are isotopic.*

In particular, if the 3-manifold M admits an $\mathbb{E}^1$-fibration over a compact surface S, this fibration is unique up to isotopy.

As a summary, if the 3-manifold M admits a Seifert fibration or an $\mathbb{E}^1$-fibration, this fibration is usually unique up to isotopy, except in a few cases which are well understood.

Having analyzed the topological aspects of these fibrations, we can now investigate the geometries corresponding to a given structure. Namely, for a fixed Seifert fibration or $\mathbb{E}^1$-fibration of M, we want to analyze the complete geometric structures on M modelled over $\mathbb{S}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \times \mathbb{E}^1$, $\mathbb{H}^2 \widetilde{\times} \mathbb{E}^1$, $\mathbb{E}^2 \widetilde{\times} \mathbb{E}^1$, $\mathbb{E}^2 \times \mathbb{E}^1$ or $\mathbb{S}^2 \widetilde{\times} \mathbb{S}^1$ for which the $\mathbb{E}^1$- or $\mathbb{S}^1$-factors give the fibration considered. We consider these geometric structures modulo the natural equivalence relation of fibration-preserving isotopy, namely we identify two such geometric structure when one is the image of the other by a fibration-preserving diffeomorphism (namely a diffeomorhism sending fiber to fiber) of M which is isotopic to the identity through a family of fibration-preserving diffeomorphisms. If M is such a fibered 3-manifold, let $\mathcal{G}_{\mathrm{f}}(M; X)$ be the moduli spaces of such fibered equivalence classes of complete geometric structures on M, with model space X and whose associated fibration is the fibration considered.

Consider the case of a Seifert fibration on M, with base 2-orbifold Σ. Let M be endowed with a geometric structure modelled over $X \times \mathbb{E}^1$ or $X \widetilde{\times} \mathbb{E}^1$, with $X = \mathbb{S}^2$, $\mathbb{E}^2$ or $\mathbb{H}^2$, and such that the Seifert fibration is defined by the $\mathbb{E}^1$-factors. To unify the notation, we write here $\mathbb{S}^2 \widetilde{\times} \mathbb{S}^1 = \mathbb{S}^2 \widetilde{\times} \mathbb{E}^1$, which is consistent since $\mathbb{S}^2 \widetilde{\times} \mathbb{S}^1$ is *locally* a twisted product of $\mathbb{S}^2$ and $\mathbb{E}^1$ although there is no globally defined space $\mathbb{S}^2 \widetilde{\times} \mathbb{E}^1$. The geometry of M then projects to a complete geometric structure over the orbifold Σ, modelled over the space X. Changing the geometric structure of M by a fibration-preserving isotopy only modifies the geometric structure of Σ by an (orbifold) isotopy. This defines a natural map from $\mathcal{G}_{\mathrm{f}}(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_{\mathrm{f}}(M; X \widetilde{\times} \mathbb{E}^1)$ to the space $\mathcal{G}(\Sigma; X)$ of isotopy classes of complete geometric structures on Σ, modelled over X.

The moduli spaces $\mathcal{G}(\Sigma; X)$ are easy to determine. One way to do this is to describe a geometric structure on Σ by gluing of elementary pieces, as in the second construction of geometric structures on surfaces of finite type in Section 1.1, and to keep track of the parameters involved. When Σ is a manifold (with no singular point), this approach goes back to Fricke and Klein, and easily extends to the framework of 2-orbifolds. In particular, this analysis is carefully worked out in Ohshika [102] for compact 2-orbifolds Σ, and the analysis easily extends to all 2-orbifolds of finite type. If the base 2-orbifold Σ of a Seifert fibration has finite type, recall that its underlying topological space $|\Sigma|$ is a surface with (possibly empty) boundary, where boundary points correspond to those points where the isotropy group is $\mathbb{Z}_2$ acting by reflection. The orbifold Σ may also have s isolated singular points, where the isotropy group is cyclic acting by rotation, c ends of 'cylindrical type', isomorphic to the manifold $\mathbb{S}^1 \times [0, \infty[$, and r ends of 'rectangular type', isomorphic to $(\mathbb{S}^1/\mathbb{Z}_2) \times [0, \infty[$ where $\mathbb{S}^1/\mathbb{Z}_2$ denotes the orbifold quotient of $\mathbb{S}^1$ by $\mathbb{Z}_2$ acting by reflection. Let $\chi(|\Sigma|)$ denote the Euler characteristic of the topological space $|\Sigma|$ underlying Σ, which should not be confused with the orbifold Euler characteristic $\chi_{\mathrm{orb}}(\Sigma)$ of Σ which we encountered in Section 2.4.

Then, when $X = \mathbb{H}^2$ and the orbifold Euler characteristic $\chi_{\mathrm{orb}}(\Sigma)$ is negative, $\mathcal{G}(\Sigma; \mathbb{H}^2)$ is homeomorphic to $\mathbb{R}^{-3\chi(|\Sigma|)-c+r+2s} \times [0, \infty[^{c+r}$, where $\chi(|\Sigma|)$, c, r and s are as above. The element of $[0, \infty[$ associated to an end of Σ which is isomorphic to $\mathbb{S}^1 \times [0, \infty[$ or $(\mathbb{S}^1/\mathbb{Z}_2) \times [0, \infty[$ is the infimum of the lengths of all 1-suborbifolds of Σ that are isotopic to $\mathbb{S}^1 \times \{0\}$ or $(\mathbb{S}^1/\mathbb{Z}_2) \times \{0\}$, respectively; this infimum is 0 exactly when the end is a cusp of finite area. In the relatively degenerate cases where the orbifold Euler characteristic of Σ is non-negative, the moduli space $\mathcal{G}(\Sigma; \mathbb{H}^2)$ is homeomorphic to the empty set, $\{0\}$, $\mathbb{R}$ or $[0, \infty[$.

Similarly, when $X = \mathbb{E}^2$, $\mathcal{G}(\Sigma; \mathbb{E}^2)$ is homeomorphic to some $\mathbb{R}^n$ with $n \leqslant 3$ or is empty. For instance, $\mathcal{G}(\Sigma; \mathbb{E}^2)$ is homeomorphic to $\mathbb{R}^3$ when Σ is the 2-torus $\mathbb{T}^2$, to $\mathbb{R}^2$ for the Klein bottle, and to $\mathbb{R}$ when Σ is the 2-orbifold whose underlying topological space is an open disk and whose singular set consists of two points with isotropy group $\mathbb{Z}_2$.

When $X = \mathbb{S}^2$, the situation is much simpler since the moduli spaces $\mathcal{G}(\Sigma; \mathbb{S}^2)$ consist of at most one point.

Again, we refer to [102] for the details of this analysis of the moduli spaces $\mathcal{G}(\Sigma; X)$ for geometric structures on base orbifolds of Seifert fibrations.

Let us return to our original problem. We have associated an element of $\mathcal{G}(\Sigma; X)$ to each element of $\mathcal{G}_{\mathrm{f}}(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_{\mathrm{f}}(M; X \mathbin{\tilde{\times}} \mathbb{E}^1)$. We also have another invariant of the geometric structures considered, namely the length l of a generic fiber of the Seifert fibration. By Theorems 2.5 and 4.1, this length l can take any positive value, except when M is compact and orientable and the Euler number $e_0 \in \mathbb{Q}$ is non-zero, in which case l is necessarily equal to $|e_0|\,\mathrm{area}(\Sigma)$.

If two geometric structures $m, m' \in \mathcal{G}_{\mathrm{f}}(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_{\mathrm{f}}(M; X \mathbin{\tilde{\times}} \mathbb{E}^1)$ have the same generic fiber length l one can easily arrange, by an isotopy of m respecting each fiber, that m and m' induce the same metric on each fiber of the Seifert fibration. The key point is that the space of oriented diffeomorphisms of the circle has the homotopy type of the circle.

One could think that, if two geometric structures $m, m' \in \mathcal{G}_{\mathrm{f}}(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_{\mathrm{f}}(M; X \mathbin{\tilde{\times}} \mathbb{E}^1)$ induce the same geometric structure on the base 2-orbifold Σ and the same metric on each fiber, then m and m' coincide. However, there is an additional invariant, which is best understood when the Seifert fibration is *oriented*, namely when we can and do choose an orientation of each fiber of the fibration which varies continuously with the fiber.

Consequently, suppose that the Seifert fibration of M is oriented, and let $m, m' \in \mathcal{G}_{\mathrm{f}}(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_{\mathrm{f}}(M; X \mathbin{\tilde{\times}} \mathbb{E}^1)$ be two geometric structures which induce the same metrics on the base Σ and on each fiber of the fibration. Note that, in this situation, the metrics m and m' coincide exactly when, at each point of M, the planes orthogonal to the fiber for m and, respectively, m' coincide. We can measure how far we are from this situation as follows. Let Σ_0 be the 2-manifold consisting of the regular points of the orbifold Σ. We then define a differential form $\omega_{m,m'} \in \Omega^1(\Sigma_0)$ of degree 1 on Σ_0 by the following property: if v is a vector in Σ_0, lift it to a vector $\widetilde{v}$ in M which is m-orthogonal to the fiber; then, $\omega_{m,m'}(v)$ is the m'-scalar product of $\widetilde{v}$ and of the unit vector tangent to the oriented fiber; it easily follows from the fact that m and m' induce the same metric on all fibers that this is independent of the choice of the lift $\widetilde{v}$. More geometrically, if α is an arc in Σ_0 and if we lift it to two arcs $\widetilde{\alpha}$ and $\widetilde{\alpha}'$ in M which are respectively m- and m'-orthogonal to the fibers and which have the same starting point, the integral of $\omega_{m,m'}$ over α is equal to the signed distance from the end point of $\widetilde{\alpha}$ to the end point of $\widetilde{\alpha}'$, for the metric induced by m and m' on the oriented fiber corresponding to the end point of α.

If α is a closed loop which is homotopic to 0 in Σ_0 and if we are considering an untwisted geometry $X \times \mathbb{E}^1$, the integral of $\omega_{m,m'}$ over α is 0 since the end points of $\widetilde{\alpha}$ and $\widetilde{\alpha}'$ are both equal to their common starting point. For a twisted geometry $X \widetilde{\times} \mathbb{E}^1$ and if α is again a closed loop homotopic to 0, we saw in Section 2.2 that the end points of $\widetilde{\alpha}$ and $\widetilde{\alpha}'$ are both obtained by shifting their starting point by an amount of $-A$, where A is the signed area enclosed by α; as a consequence, these end points coincide and it again follows that the integral of $\omega_{m,m'}$ over α is equal to 0. In both cases, it follows that the differential form $\omega_{m,m'}$ is closed. In particular, it defines a cohomology class in $H^1(\Sigma_0; \mathbb{R})$. The integral of $\omega_{m,m'}$ over a loop in Σ_0 which goes around an isolated singular point of Σ is trivial; it follows that this element of $H^1(\Sigma_0; \mathbb{R})$ comes from a cohomology class of $H^1(|\Sigma|; \mathbb{R})$, where $|\Sigma|$ denotes the topological space underlying the orbifold Σ.

If m'' is another metric which is isotopic to m' by an isotopy respecting each fiber and which induces the same metric as m' on each fiber, one easily sees that $\omega_{m,m''} = \omega_{m,m'} + df$ where the function $f : \Sigma_0 \to \mathbb{R}$ measures the amount of rotation of the isotopy on each fiber. It follows that the element of $H^1(|\Sigma|; \mathbb{R})$ that is represented by $\omega_{m,m'}$ depends only on the classes of m and m' in $\mathcal{G}_f(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_f(M; X \widetilde{\times} \mathbb{E}^1)$.

The metric m' can easily be recovered from the closed differential form $\omega_{m,m'}$. Therefore, the space of elements $m' \in \mathcal{G}_f(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_f(M; X \widetilde{\times} \mathbb{E}^1)$ which have the same generic fiber length and the same image in $\mathcal{G}(\Sigma; X)$ as m is naturally identified to $H^1(|\Sigma|; \mathbb{R})$.

This analysis works when the Seifert fibration of M is oriented. However, it easily extends to the general case, provided we replace $H^1(|\Sigma|; \mathbb{R})$ by the cohomology group $H^1(|\Sigma|; \widehat{\mathbb{R}})$ with coefficients twisted by the orientation cocycle of the Seifert fibration. More precisely, consider the space $\widehat{M}$ of pairs (x, o) where $x \in M$ and o is a local orientation of the fiber of the Seifert fibration at x, with the natural topology. This manifold $\widehat{M}$ is a 2-fold covering of M, and the Seifert fibration of M lifts to a Seifert fibration of $\widehat{M}$ which is canonically oriented by choosing the orientation o at each $(x, o) \in \widehat{M}$. The 2-fold covering $\widehat{M} \to M$ descends to a 2-fold covering $|\widehat{\Sigma}| \to |\Sigma|$ between the spaces underlying their base 2-orbifolds. The twisted cohomology group $H^1(|\Sigma|; \widehat{\mathbb{R}})$ is defined by consideration of cochains on $|\widehat{\Sigma}|$ which are anti-equivariant with respect to the covering automorphism $\tau : |\widehat{\Sigma}| \to |\widehat{\Sigma}|$ that exchanges the two sheets of the covering $|\widehat{\Sigma}| \to |\Sigma|$, namely of cochains c such that $\tau^*(c) = -c$. As indicated, the extension of the above analysis to this twisted context is automatic.

A careful consideration of the argument actually shows:

THEOREM 5.4. *Let the* 3*-manifold M be endowed with a Seifert fibration with base* 2*-orbifold Σ. For $X = \mathbb{S}^2$, $\mathbb{E}^2$ or $\mathbb{H}^2$, let $\mathcal{G}_f(M; X \times \mathbb{E}^1)$ and $\mathcal{G}_f(M; X \widetilde{\times} \mathbb{E}^1)$ be the space of complete geometric structures modelled over the spaces indicated, where the $\mathbb{E}^1$-factors correspond to the fibers of the Seifert fibration, where these geometric structures are considered up to fibration-preserving isotopy. Let $\mathcal{G}(\Sigma; X)$ denote the space of isotopy classes of complete geometric structures on the orbifold Σ modelled over X. Then, if the spaces $\mathcal{G}_f(M; X \times \mathbb{E}^1)$ or $\mathcal{G}_f(M; X \widetilde{\times} \mathbb{E}^1)$ are non-empty (and compare Theorems* 2.5(i) *and* 4.1 *for this), the natural maps $\mathcal{G}_f(M; X \times \mathbb{E}^1) \to \mathcal{G}(\Sigma; X)$ and $\mathcal{G}_f(M; X \widetilde{\times} \mathbb{E}^1) \to \mathcal{G}(\Sigma; X)$*

are trivial bundles with fiber $H^1(|\Sigma|; \widehat{\mathbb{R}})$ *or* $H^1(|\Sigma|; \widehat{\mathbb{R}}) \times]0, \infty[$, *where* $H^1(|\Sigma|; \widehat{\mathbb{R}})$ *is the twisted cohomology group defined above, and where the factor* $]0, \infty[$ *corresponds to the length of the generic fiber and occurs in all cases unless when* M *is compact orientable and the Seifert fibration has non-trivial Euler number* $e_0 \in \mathbb{Q}$.

For interval bundles, the same argument proves:

THEOREM 5.5. *Let the* 3*-manifold* M *be endowed with an (open) interval bundle structure with base surface* S. *For* $X = \mathbb{S}^2$, $\mathbb{E}^2$ *or* $\mathbb{H}^2$, *let* $\mathcal{G}_f(M; X \times \mathbb{E}^1)$ *be the space of complete geometric structures modelled over the spaces indicated, where the* $\mathbb{E}^1$*-factors correspond to the fibers of the interval bundle, where these geometric structures are considered up to fibration-preserving isotopy. Let* $\mathcal{G}(\Sigma; X)$ *denote the space of isotopy classes of complete geometric structures on the orbifold* Σ *modelled over* X. *Then, the natural map* $\mathcal{G}_f(M; X \times \mathbb{E}^1) \to \mathcal{G}(\Sigma; X)$ *is a trivial bundle with fiber* $H^1(|\Sigma|; \widehat{\mathbb{R}})$, *where* $H^1(|\Sigma|; \widehat{\mathbb{R}})$ *is the twisted cohomology group defined above.*

6. Applications of 3-dimensional geometric structures

We conclude with a discussion of a few purely topological applications of the use of geometric (mostly, hyperbolic) structures on 3-manifolds. This selection is only intended to give a sample of such applications. It clearly reflects the personal tastes of the author, and is by no means intended to be exhaustive.

6.1. *Knot theory*

The area where the use of geometric structures, essentially hyperbolic geometry, has had the greatest practical impact is probably knot theory. Let L be a *link* in the 3-sphere $\mathbb{S}^3$, meaning that L is a 1-dimensional submanifold of $\mathbb{S}^3$. A connected link is also called a *knot*. Knot theory aims at classifying all such links up to diffeomorphism of $\mathbb{S}^3$; see for instance the standard references [116,23]. To show that two links are different modulo diffeomorphism of $\mathbb{S}^3$, the traditional method is to use algebraic topology to extract some algebraic invariants of these links; if the invariants computed happen to be different, this shows that the links are different.

The consideration of hyperbolic structures provides a completely new type of invariants. Indeed, the Hyperbolization Theorem 4.3 shows that the complement $\mathbb{S}^3 - L$ of a link $L \subset \mathbb{S}^3$ admits a finite volume complete hyperbolic structure unless one of the following holds:

(i) $\mathbb{S}^3 - L$ contains an embedded essential 2-torus.

(ii) $\mathbb{S}^3 - L$ admits a Seifert fibration.

In case (ii), the Seifert fibration of $\mathbb{S}^3 - L$ can be chosen so that it extends to a Seifert fibration of $\mathbb{S}^3$, for which L consists of finitely many fibers of this fibration. Such links are called *torus links*. Since the Seifert fibrations $\mathbb{S}^3$ are easily classified [128,104], torus links are easily classified.

In case (i), L is said to be a *satellite link*. The Characteristic Torus Decomposition of Theorem 3.4 provides a canonical factorization of a satellite link into links which are, either torus links, or non-satellite links.

This reduces the analysis of all links to those whose complement admits a finite volume complete hyperbolic structure. As observed in Section 5.1, Mostow's Rigidity Theorem 5.1 implies that any geometric invariant of this hyperbolic structure is a topological invariant of the link complement $\mathbb{S}^3 - L$, and therefore of the link $L \subset \mathbb{S}^3$. Among such geometric invariants, we already mentioned the volume of the hyperbolic structure. A more powerful invariant of the hyperbolic structure is its Ford domain, which we now briefly describe; see for instance Maskit's book [83, Chapter IV.F] for details.

Let M be a non-compact orientable hyperbolic 3-manifold of finite volume. We saw in Theorem 2.9 and in our discussion of Theorem 4.5 that M has finitely many ends and that each end e has a neighborhood U_e isometric to a model H_e/Γ_e, where $H_e \subset \mathbb{H}^3$ is a horoball $\{(u, v, w) \in \mathbb{R}^3;\ w \geqslant w_e\}$ and where the group $\Gamma_e \cong \mathbb{Z}^2$ acts by horizontal translations. Adjusting the constants $w_e > 0$, and in particular choosing them large enough, we can arrange that these neighborhoods U_e are pairwise disjoint and have the same volume. For every $x \in M$, consider those arcs which join x to the union of the neighborhoods U_e and, among those arcs, consider those which are shortest. The *Ford domain* of M consists of those x for which there is a unique such shortest arc joining x to the U_e. One easily sees that the Ford domain is independent of the cusp neighborhoods U_e, provided they are chosen sufficiently small and of equal volumes.

To each end e of M is associated a component of the Ford domain, consisting of those points which are closer to U_e than to any other $U_{e'}$. This component is isometric to $\mathrm{int}(P_e)/\Gamma_e$ where P_e is a locally finite convex polyhedron in $\mathbb{H}^3$ which is invariant under the horizontal translation group $\Gamma_e \cong \mathbb{Z}^2$. Here, a *locally finite convex polyhedron* in $\mathbb{H}^3$ is the intersection P of a family of closed half-spaces bounded by totally geodesic planes in $\mathbb{H}^3$, such that every point of P has a neighborhood which meets only finitely many of the boundaries of these half-spaces. The reader should beware of a competing terminology, used by many authors, where the Ford domain is defined as the collection of the polyhedra P_e.

By construction, the polyhedra P_e, endowed with the action of the group $\Gamma_e \cong \mathbb{Z}^2$, are uniquely determined modulo isometry of $\mathbb{H}^3 = \{(u, v, w) \in \mathbb{R}^3;\ w > 0\}$ respecting ∞, namely modulo homothety and euclidean isometry of $\mathbb{R}^3$ respecting $\mathbb{H}^3$. In particular these polyhedra P_e, endowed with their action of Γ_e, are geometric invariants of the hyperbolic metric of M, and therefore are topological invariants of M by Mostow's Rigidity Theorem 5.1. In particular, an invariant extracted from the Ford domain is an invariant of M; simple examples include the number of vertices, edges and faces of $\partial P_e/\Gamma_e$, or the way these faces fit together (namely the combinatorial structure of the polyhedral decomposition of $\partial P_e/\Gamma_e$), or the geometry of these faces.

The Ford domain is such a powerful invariant that it is possible to reconstruct M from it. Indeed, it comes equipped with an isometric pairing of its faces. The manifold M is then obtained from the disjoint union of P_e/Γ_e by gluing its faces through this pairing. Actually, in the case where M is the complement $\mathbb{S}^3 - L$ of a link L, reconstructing M from the Ford domain is not sufficient to characterize the link L, since there are different links which have homeomorphic complements; see for instance [116, Section 9.H]. However, L is completely determined if we specify the meridians of the components of L, an information which is easily encoded in the groups Γ_e associated to the ends of M; again, see [116, Chapter 9]. For knots, this meridian information is in fact unnecessary by a deep theorem of Gordon and Luecke [44].

The problem is of course to be able to compute this invariant in practice. The first problem is that the Hyperbolization Theorem 4.3 is only an abstract existence theorem, and that the proofs available are non-constructive. The second problem is that, even if we are given a hyperbolic structure on M, for instance under the form of a free isometric properly discontinuous action of a group Γ on $\mathbb{H}^3$ such that $\mathbb{H}^3/\Gamma$ is diffeomorphic to M, it may be hard to explicitly determine the corresponding Ford domain.

The pioneering work in this area was developed by Riley [114,115]. For certain links L in $\mathbb{S}^3$, he used a computer to find finitely many isometries $A_1, \ldots, A_n$ of $\mathbb{H}^3$ such that the group Γ generated by the A_i acts freely and properly discontinuously on $\mathbb{H}^3$ and such that Γ is abstractly isomorphic to the fundamental group of $\mathbb{S}^3 - L$; Waldhausen's Theorem 3.11 on homotopy equivalences of Haken 3-manifolds then guarantees that M is diffeomorphic to $\mathbb{H}^3/\Gamma$. Riley also determined the Ford domains of these hyperbolic manifolds. However, the use of a computer raises the question of rounding errors: For instance, if we compute the isometry corresponding to a word in the A_i and if the value provided by the computer is the identity, does this mean that this isometry is really the identity (which is what we need to make sure that the algebraic structure of Γ is the one expected), or does this just mean that this isometry is very close to the identity (which could have dire consequences for the proper discontinuity of the action)? In these examples, once the first set of computations by the computer had provided him with appropriate conjectures on what the generators A_i and the shape of the Ford domain should be, Riley was able to justify these computations *a posteriori* by exact arithmetic computations in a number field. Namely, he then rigorously proved that the A_i provided by the computer could be approximated by isometries generating a group Γ with the required properties, and that the Ford domain determined by the computer was indeed an approximation of the exact Ford domain of Γ.

Riley's group theoretic approach is unfortunately not very efficient from a computational point of view. The software SnapPea [153], later developed by J. Weeks (with collaborators for some additional features), uses a more geometric approach and works incredibly well in practice. Given a link L in $\mathbb{S}^3$, SnapPea computes a hyperbolic structure on the complement $\mathbb{S}^3 - L$, if it exists, and describes its Ford domain. SnapPea also computes various invariants of this hyperbolic structure, as well as hyperbolic structures on 3-manifolds obtained by 'sufficiently complicated' Dehn filling, as in the Hyperbolic Dehn Filling Theorem 4.4.

In practice, SnapPea works very fast for links with a reasonable number of crossings. However, there is a drawback which has to do with rounding errors. Since SnapPea works only with finite precision, its outputs can mathematically only be considered as conjectural approximations to the exact situation. In theory, it is possible to justify these guesses *a posteriori* by using exact arithmetic as in [114,115], but this is often not workable in practice. In any case, ever since the first versions of SnapPea started circulating, it has established itself as an invaluable tool to study examples, and make and disprove conjectures, in hyperbolic geometry and knot theory. In particular, it has been extensively used to establish useful tables of links and hyperbolic 3-manifolds, with the *caveat* about the theoretical reliability of the output due to rounding errors; see for instance [1,54,152].

There is something interesting about the algorithm used by SnapPea to find a hyperbolic structure on a link complement. It is a variation of the famous method used by Thurston in [138] to construct a hyperbolic structure on the complement of the figure eight knot. Namely it decomposes the link complement into finitely many 'ideal simplices', with all vertices at infinity, and tries to put a hyperbolic metric on each of these ideal simplices, so that the metrics fit nicely along the faces and edges of the decomposition. When SnapPea fails to find such hyperbolic structures on the ideal simplices, it uses various combinatorial schemes to modify the decomposition into ideal simplices until it reaches a solution. What is remarkable is that, although this algorithm works extremely well in practice, there is, at this point, no general proof of the Hyperbolization Theorem 4.3 for link complements which is based on this strategy. Conversely, the proofs of Theorem 4.3 mentioned in Section 4.3 are usually non-constructive.

6.2. *Symmetries of 3-manifolds*

One of the early successes in the use of hyperbolic geometry to study the topology of 3-manifolds was the proof of the following conjecture of P.A. Smith.

THEOREM 6.1 (Smith (ex-)Conjecture). *Let $f:\mathbb{S}^3 \to \mathbb{S}^3$ be an orientation-preserving periodic diffeomorphism of the 3-sphere whose fixed point set is non-empty. Then f is conjugate to a rotation of $\mathbb{S}^3 = \mathbb{R}^3 \cup \{\infty\}$ by a diffeomorphism of $\mathbb{S}^3$.*

The original proof, expounded in [10], is a combination of various ingredients, coming from different branches of mathematics. The main idea is to consider the fixed point set L of f, which is a knot in $\mathbb{S}^3$ (the connectedness of L was Smith's original result in [133]). Various minimal surface arguments reduce the problem to the case where the complement $\mathbb{S}^3 - L$ contains no f-invariant essential surface. If there is no such essential surface, the Hyperbolization Theorem 4.3 provides either a Seifert fibration or a finite volume hyperbolic structure on $\mathbb{S}^3 - L$. An easy fundamental group computation in the case of a Seifert fibration, and the use of much more subtle algebraic and number theoretic properties of subgroups of $\mathrm{PSL}_2(\mathbb{C})$ in the case of a hyperbolic structure, then enable one to complete the proof.

This original proof of the Smith Conjecture is now superseded by the Orbifold Geometrization Theorem 4.7.

Other topological applications involve the *symmetry group* of a 3-manifold M with (possibly empty) boundary, defined as the group $\pi_0\mathrm{Diff}(M)$ of isotopy classes of diffeomorphisms of M.

If M is endowed with complete hyperbolic metric with finite volume and with totally geodesic boundary, Mostow's Rigidity Theorem 5.1 says that every diffeomorphism f of M is homotopic to an isometry of the metric. Theorem 3.12 for the case when M is Haken, and Theorem 6.11 of the next section for the general case, show that f is actually isotopic to an isometry. In addition, it is not hard to see that two distinct isometries of M cannot be homotopic. This proves that, if $\mathrm{Isom}(M)$ is the isometry group of M, the natural map from $\mathrm{Isom}(M)$ to $\pi_0\,\mathrm{Diff}(M)$ is a bijection.

THEOREM 6.2. *Let the* 3*-manifold* M *admit a complete hyperbolic structure with finite volume and with totally geodesic boundary* (*compare Theorems* 2.14 *and* 4.3). *Then, there is a finite group* G *acting on* M *such that the natural map from* G *to* $\pi_0 \operatorname{Diff}(M)$ *is a bijection. In particular, the group* $\pi_0 \operatorname{Diff}(M)$ *is finite.*

This improves a result of K. Johannson, who had proved in [64] that $\pi_0\operatorname{Diff}(M)$ is finite when the compact 3-manifold M with boundary is Haken and contains no essential disk, 2-torus or annulus. When M contains an embedded essential disk, 2-torus or annulus, the consideration of Dehn twists along this surface usually implies that $\pi_0 \operatorname{Diff}(M)$ is infinite.

In Theorem 6.2, the fact that $\pi_0 \operatorname{Diff}(M)$ can be realized by the action of the finite group is a powerful tool. See [16] or [39] for a few applications to problems in classical topology.

Another important property comes from the Orbifold Geometrization Theorem 4.7. Let the 3-manifold M admit a finite volume hyperbolic structure, and let G be a finite group acting on M. We can then consider the 3-orbifold M/G. Suppose that the fixed point set of some non-trivial element of G has dimension at least 1, namely that the singular set of the orbifold M/G is at least 1-dimensional.

The orbifold M/G then satisfies the hypotheses of the Orbifold Geometrization Theorem 4.7. Indeed, every 2-suborbifold of M/G is uniformized by its pre-image in M, and is therefore uniformizable. An essential sphere or torus 2-suborbifold would lift to a 2-sphere, projective plane, 2-torus or Klein bottle in M, which would have to be essential by the fact that all finite group actions on $\mathbb{B}^3$ are standard [91] or by the Equivariant Dehn Lemma [89,90]; but this would contradict the existence of the hyperbolic structure of M, by Theorem 2.9. See for instance [21, Section D] for details.

At this point, Theorem 4.7 asserts that the orbifold M/G admits a finite volume geometric structure. This geometric structure of M/G lifts to a G-invariant geometric structure on the manifold M. Since we already know that M admits a finite volume hyperbolic structure, this G-invariant geometric structure is necessarily hyperbolic by Theorem 2.10. In addition, Mostow's Rigidity Theorem 5.1, together with Theorems 3.12 and 6.11, show that this G-invariant hyperbolic structure is isotopic to the original one. If we use the isotopy to conjugate the group action instead of changing one hyperbolic metric to the other, this proves:

THEOREM 6.3. *Let the finite group* G *act on a* 3*-manifold* M *which admits a finite volume hyperbolic structure with totally geodesic boundary. Then the action is conjugate to an isometric action by a diffeomorphism of* M *which is isotopic to the identity.*

Actually, we only discussed the case without boundary. But, as usual, the case with boundary is easily deduced from this one by consideration of the double manifold DM with the action of the group $G \oplus \mathbb{Z}/2$, where the factor $\mathbb{Z}/2$ acts by exchange of the two halves of DM.

6.3. *Covering properties*

A classical problem in geometric topology is to decide if a non-compact manifold has finite topological type. A particularly interesting source of non-compact manifolds is provided by (connected) coverings $\widetilde{M} \to M$, where the manifold M has finite topological type. It turns out that the use of hyperbolic geometry can provide some answers to this type of problems. We give here a few samples of the type of results which can be obtained through this approach.

An immediate corollary of the existence theorems for hyperbolic structures is the following:

THEOREM 6.4. *Let M be a 3-manifold which admits a complete hyperbolic structure or, more generally, a complete metric of non-positive curvature (compare Theorems 4.3, 4.4, 4.5 or 4.7). Then the universal cover of M is homeomorphic to the euclidean space $\mathbb{E}^3$. More generally, for every connected covering $\widetilde{M} \to M$ where the fundamental group $\pi_1(\widetilde{M})$ is abelian, $\widetilde{M}$ has finite topological type.*

The first statement is a celebrated theorem of Hadamard (see for instance do Carmo [26, Chapter 7] or Eberlein [33, Section 1.4], and compare Sections 1.2 and 2.1). The second one follows from the fact that isometric actions of abelian groups on a simply connected manifold of non-positive curvature are easily classified; see [33, Section 1.9].

A more subtle geometric argument leads to the following purely topological result.

THEOREM 6.5. *Let M be the interior of a compact 3-manifold $\overline{M}$ with boundary which contains no essential 2-sphere, projective plane, 2-torus or annulus, and such that at least one component of $\partial\overline{M}$ has negative Euler characteristic. If $\widetilde{M}$ is a connected covering of M whose fundamental group is finitely generated, then $\widetilde{M}$ has finite topological type.*

A proof can be found in Morgan [98, Proposition 7.1]. The idea is to use the Hyperbolization Theorem 4.3 to endow M with a complete hyperbolic metric. The metric provided by the proof of Theorem 4.3 is of a certain type, called "geometrically finite". A relatively simple observation then shows that the lift of this metric to $\widetilde{M}$ is also geometrically finite (this is where we need the hypothesis that at least one component of $\partial\overline{M}$ has negative Euler characteristic), from which it follows that $\widetilde{M}$ has finite topological type. Presumably, the hypothesis that $\overline{M}$ contains no essential 2-sphere, projective plane, 2-torus or annulus is unnecessary, as the general case should follow from the above one and from the use of the characteristic splittings of Section 3.

Similar results can be obtained from relatively deep results on the geometry of ends of hyperbolic 3-manifolds, obtained by Thurston [139] and the author [18]. Indeed, one of the topological consequences of this analysis is the following result:

THEOREM 6.6. *Let the 3-manifold M admit a complete hyperbolic metric. Assume that the fundamental group $\pi_1(M)$ is finitely generated and does not (algebraically) split as a non-trivial free product of two groups. Then M has finite topological type.*

This immediately gives the following corollary:

COROLLARY 6.7. *Let the 3-manifold M admit a complete hyperbolic metric (compare Theorems 4.3, 4.4, 4.5 or 4.7), and consider a covering $\widetilde{M} \to M$ with $\widetilde{M}$ connected. If* *the fundamental group $\pi_1(\widetilde{M})$ is finitely generated and does not split as a non-trivial free product, then $\widetilde{M}$ has finite topological type.*

We should also include in this section the important residual finiteness property for fundamental groups of 3-manifolds.

Recall that a group G is *residually finite* if, for every non-trivial $g \in G$, there exists a finite index subgroup of G which does not contain g. This property has important algebraic consequences for the group G; see for instance Magnus [80]. However, what is more of interest to topologists is that, when G is the fundamental group $\pi_1(M)$ of a manifold M, the residual finiteness of $\pi_1(M)$ is equivalent to the following topological property: For every compact subset K of the universal covering $\widetilde{M}$, there is a connected *finite index* covering $\widetilde{M}_0$ of M such that the natural projection $\widetilde{M} \to \widetilde{M}_0$ is injective on K. The equivalence between these two properties is an easy exercise, and the topological property is very useful in practice.

The connection with geometric structures is a theorem of Mal'cev [82] which asserts that every finitely generated group of matrices (with entries in an arbitrary commutative field) is residually finite. Since all isometry groups of the 3-dimensional geometries embed in matrix groups, a geometric structure on the 3-manifold M embeds $\pi_1(M)$ in such a matrix group. This proves:

THEOREM 6.8. *Let the finite type 3-manifold M admit a geometric structure. Then its fundamental group $\pi_1(M)$ is residually finite.*

Let M be a compact Haken 3-manifold. For the characteristic torus decomposition T of Section 3.2, the geometrization results of Sections 4.2 and 4.3 provide geometric structures on the components of $M - T \cup \partial M$, and in particular show that the fundamental groups of these components are residually finite. It is non-trivial to conclude from this that the fundamental group of M itself is residually finite, but this is indeed a result of Hempel [52]:

THEOREM 6.9. *Let M be a Haken 3-manifold. Then the fundamental group $\pi_1(M)$ is residually finite.*

Note that residual finiteness is preserved under finite index extensions. Therefore, Theorem 6.9 holds under the slightly weaker hypothesis that the manifold M is *virtually Haken*, namely admits a finite cover which is Haken.

6.4. *Topological rigidity of hyperbolic 3-manifolds*

Finally, we mention some recent results which prove for hyperbolic 3-manifolds the results obtained by Waldhausen for homotopy equivalences and isotopies of Haken manifolds, as discussed in Section 3.5.

THEOREM 6.10. *Let $f : M \to N$ be a homotopy equivalence between a compact hyperbolic 3-manifold M and a 3-manifold N which contains no essential 2-sphere. Then f is homotopic to a diffeomorphism.*

The requirement that the 3-manifold N contains no essential 2-sphere is here only to circumvent any possible counter-example to the Poincaré conjecture. Indeed, it easily follows from the hyperbolic structure of M that every decomposition of the fundamental groups $\pi_1(M) \cong \pi_1(N)$ as a free product must be trivial. Therefore, the hypothesis that N is homotopy equivalent to M already implies that every 2-sphere embedded in N must bound in N a homotopy 3-ball, namely a contractible 3-submanifold of N.

THEOREM 6.11. *Let f_0, $f_1 : M \to M$ be two diffeomorphisms of a compact hyperbolic 3-manifold M. If f_0 and f_1 are homotopic, then they are isotopic.*

Theorems 6.10 and 6.11 were proved in three steps. First, by an elegant but comparatively simple argument, Gabai proved in [41] that the conclusion of these theorems holds in finite covers of M and N; the main ingredients of this part of the proof are residual finiteness and the techniques developed by Waldhausen in [145]. Then, by more elaborate arguments which use in a crucial way the geometry at infinity of hyperbolic 3-space, Gabai was able to prove Theorems 6.10 and 6.11 under the additional assumption that the hyperbolic 3-manifold M satisfies a certain "insulator condition" [42]. He also conjectured that any compact hyperbolic 3-manifold satisfies this insulator condition. This easily translates to a similar conjecture for discrete 2-generator subgroups of the isometry group $\mathrm{Isom}^+(\mathbb{H}^3) = \mathrm{PSL}_2(\mathbb{C})$. Now, a 2-generator subgroup of $\mathrm{PSL}_2(\mathbb{C})$ is determined by its generators, and therefore by a finite number of complex parameters. It quickly became clear that any counterexample to the conjecture would provide one in a certain compact portion of the corresponding parameter space. Gabai, R. Meyerhoff and N. Thurston then scanned this compact region of the parameter space, and, through a careful control of rounding errors, were able to rigorously prove that every compact hyperbolic 3-manifold M satisfies the insulator condition. The details of this computer-assisted part of the proof of Theorems 6.10 and 6.11 can be found in [43].

References

[1] C.C. Adams, M. Hildebrand and J. Weeks, *Hyperbolic invariants of knots and links*, Trans. Amer. Math. Soc. **326** (1991), 1–56.

[2] L.V. Ahlfors and L. Sario, *Riemann Surfaces*, Princeton University Press (1960).

[3] J.W. Alexander, *On the deformation of an n-cell*, Proc. Nat. Acad. Sci. USA **9** (1923), 406–407.

[4] J.W. Alexander, *On the subdivision of 3-space by a polyhedron*, Proc. Nat. Acad. Sci. USA **10** (1924), 6–8.

[5] M.T. Anderson, *Scalar curvature and geometrization conjectures for 3-manifolds*, Comparison Geometry, M.S.R.I. Publications, Vol. 30, Cambridge University Press (1997), 47–80.

[6] E.M. Andreev, *On convex polyhedra in Lobačevskiĭ spaces* (Russian), Mat. Sbornik **81** (123) (1970), 445–478; English translation: Math. USSR Sb. **10** (1970), 413–440.

[7] E.M. Andreev, *On convex polyhedra of finite volume in Lobačevskiĭ spaces* (Russian), Mat. Sbornik **83** (125) (1970), 256–260; English translation: Math. USSR Sb. **12** (1970), 255–259.

[8] K. Asano, *Homeomorphisms of prism manifolds*, Yokohama Math. J. **26** (1978), 19–25.

[9] L. Auslander, *An account of the theory of crystallographic groups*, Proc. Amer. Math. Soc. **16** (1965), 1230–1236.

[10] H. Bass and J.W. Morgan (eds), *The Smith Conjecture*, Academic Press, New York (1984).

[11] R. Benedetti and C. Petronio, *Lectures on Hyperbolic Geometry*, Universitext Series, Springer, Berlin (1992).
[12] L. Bieberbach, *Über die Bewegensgruppen des Euklidischen Räume* I, Math. Ann. **70** (1910), 297–336; II, Math. Ann. **72** (1912), 400–412.
[13] S. Bleiler and C.D. Hodgson, *Spherical space forms and Dehn filling*, Topology **35** (1996), 809–833.
[14] M. Boileau and J.-P. Otal, *Scindements de Heegaard et groupe des homéotopies des petites variétés de Seifert*, Invent. Math. **106** (1991), 85–107.
[15] M. Boileau and J. Porti, *Geometrization of 3-orbifolds of Cyclic Type*, Astérisque, Société Mathématique de France, to appear.
[16] F. Bonahon, *Cobordism of automorphisms of surfaces*, Ann. Éc. Norm. Sup. **16** (1983), 237–270.
[17] F. Bonahon, *Difféotopies des espaces lenticulaires*, Topology **22** (1983), 305–314.
[18] F. Bonahon, *Bouts des variétés hyperboliques de dimension* 3, Ann. of Math. **124** (1986), 71–158.
[19] F. Bonahon and J.-P. Otal, *Variétés hyperboliques à géodésiques arbitrairement courtes*, Bull. London Math. Soc. **20** (1988), 255–261.
[20] F. Bonahon and L.C. Siebenmann, *The classification of Seifert fibred 3-orbifolds*, Low Dimensional Topology, London Math. Soc. Lecture Notes Series, Vol. 95, R. Fenn, ed., Cambridge University Press (1985), 19–85.
[21] F. Bonahon and L.C. Siebenmann, *The characteristic toric splitting of irreducible compact 3-orbifolds*, Math. Ann. **278** (1987), 441–479.
[22] S. Boyer, *Dehn surgery on knots*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 165–218.
[23] G. Burde and H. Zieschang, *Knots*, de Gruyter (1985).
[24] R.D. Canary, *Ends of hyperbolic 3-manifolds*, J. Amer. Math. Soc. **6** (1993), 1–35.
[25] J.W. Cannon and C.D. Feustel, *Essential embeddings of annuli and Möbius bands in 3-manifolds*, Trans. Amer. Math. Soc. **215** (1976), 219–239.
[26] M.P. do Carmo, *Riemannian Geometry*, Birkhäuser (1992).
[27] A. Casson and D. Jungreis, *Convergence groups and Seifert fibered 3-manifolds*, Invent. Math. **118** (1994), 441–456.
[28] T. Comar, *Hyperbolic Dehn surgery and convergence of Kleinian groups*, Doctoral thesis, University of Michigan (1996).
[29] D. Cooper, C.D. Hodgson and S.P. Kerckhoff, *Three-dimensional orbifolds and cone manifolds (with a postface by S. Kojima)*, Mem. Math. Soc. Japan, Vol. 5 (2000).
[30] W.D. Dunbar, *Classification of solvorbifolds in dimension three*, Braids (Santa Cruz, 1986), Contemp. Math., Vol. 78, Amer. Math. Soc., Providence, RI (1988), 207–216.
[31] W.D. Dunbar, *Nonfibering spherical 3-orbifolds*, Trans. Amer. Math. Soc. **341** (1994), 121–142.
[32] P. Du Val, *Homographies, Quaternions and Rotations*, Clarendon Press (1964).
[33] P.B. Eberlein, *Geometry of Nonpositively Curved Manifolds*, Univ. of Chicago Press (1996).
[34] D.B.A. Epstein, *Projective planes in 3-manifolds*, Proc. London Math. Soc. **11** (1961), 469–484.
[35] D.B.A. Epstein, *Curves on 2-manifolds and isotopies*, Acta Math. **115** (1966), 83–107.
[36] D.B.A. Epstein and R.C. Penner, *Euclidean decompositions of non-compact hyperbolic manifolds*, J. Differential Geom. **27** (1988), 67–80.
[37] A. Fathi, F. Laudenbach and V. Poenaru, *Travaux de Thurston sur les Surfaces*, Astérisque, Vols 66–67, Société Mathématique de France (1979).
[38] C.D. Feustel, *On the torus theorem for 3-manifolds*, Trans. Amer. Math. Soc. **217** (1976), 45–57.
[39] E. Flapan, *Rigidity of graph symmetries in the 3-sphere*, J. Knot Theory Ramifications **4** (1995), 373–388.
[40] D. Gabai, *Convergence groups are Fuchsian groups*, Ann. of Math. **136** (1992), 447–510.
[41] D. Gabai, *Homotopy hyperbolic 3-manifolds are virtually hyperbolic*, J. Amer. Math. Soc. **7** (1994), 193–198.
[42] D. Gabai, *On the geometric and topological rigidity of hyperbolic 3-manifolds*, J. Amer. Math Soc. **10** (1997), 37–74.
[43] D. Gabai, R. Meyerhoff and N. Thurston, *Homotopy hyperbolic 3-manifolds are hyperbolic*, Ann. of Math., to appear.
[44] C. McA. Gordon and J. Luecke, *Knots are determined by their complements*, J. Amer. Math. Soc. **2** (1989), 371–415.

[45] E. Goursat, *Sur les substitutions orthogonales et les divisions régulières de l'espace*, Ann. Scient. Éc. Norm. Sup. **6** (1889), 2–102.

[46] A. Gramain, *Le groupe d'homotopie du groupe des diffeomorphismes d'une surface compacte*, Ann. Scient. Éc. Norm. Sup. **6** (1973), 53–66.

[47] M. Gromov and W.P. Thurston, *Pinching constants for hyperbolic manifolds*, Invent. Math. **89** (1987), 1–12.

[48] A. Hatcher, *Homeomorphisms of sufficiently large P^2-irreducible 3-manifolds*, Topology **5** (1976), 343–347.

[49] W. Heil, *On P^2-irreducible 3-manifolds*, Bull. Amer. Math. Soc. **75** (1969), 772–775.

[50] W. Heil, *On 3-manifolds with sufficiently large decomposition*, Canad. J. Math. **23** (1971), 746–748; Corrigendum: Canad. J. Math. **25** (1973), 1120.

[51] J. Hempel, *3-Manifolds*, Ann. Math. Studies 86, Princeton University Press (1976).

[52] J. Hempel, *Residual finiteness for 3-manifolds*, Combinatorial Group Theory and Topology, Ann. Math. Studies, Vol. 111, S.M. Gersten and J.R. Stallings, eds, Princeton University Press (1987), 379–396.

[53] H. Hendriks and F. Laudenbach, *Difféomorphismes des sommes connexes en dimension trois*, Topology **23** (1984), 423–443.

[54] S.R. Henry and J.R. Weeks, *Symmetry groups of hyperbolic knots and links*, J. Knot Theory Ramifications **1** (1992), 185–201.

[55] C.D. Hodgson, *Geometric structures on 3-dimensional orbifolds: Notes on Thurston's proof*, Lectures Notes, Columbia University (1987).

[56] C.D. Hodgson and J.H. Rubinstein, *Involutions and isotopies of lens spaces*, Knot Theory and Manifolds (Vancouver, BC, 1983), Lecture Notes in Math., Vol. 1144, Springer, Berlin (1985), 60–96.

[57] N.V. Ivanov, *Homotopy of spaces of diffeomorphisms of some three-dimensional manifolds*, Studies in Topology IV, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **122** (1982), 72–103, 164–165 (Russian).

[58] W. Jaco, *Lectures on Three-Manifold Topology*, C.B.M.S. Regional Conference Series in Mathematics, Vol. 43, Amer. Math. Soc., Providence, RI (1980).

[59] W. Jaco and P.B. Shalen, *A new decomposition theorem for irreducible sufficiently-large 3-manifolds*, Proceedings of Symposia in Pure Mathematics, Vol. 32, Part 2, R.J. Milgram, ed., Amer. Math. Soc., Providence, RI (1978), 71–84.

[60] W. Jaco and P.B. Shalen, *Seifert Fibered Spaces in 3-Manifolds*, Memoirs Amer. Math. Soc., Vol. 220, Amer. Math. Soc., Providence, RI (1979).

[61] T. Janssen, *Crystallographic Groups*, North-Holland, Amsterdam (1973).

[62] M. Jenkins and W.D. Neumann, *Lectures on Seifert Manifolds*, Brandeis Lecture Notes, Vol. 2, Brandeis University (1983).

[63] K. Johannson, *Homotopy Equivalences of 3-Manifolds with Boundary*, Lecture Notes in Math., Vol. 761, Springer, Berlin (1979).

[64] K. Johannson, *On the mapping class group of simple 3-manifolds*, Topology of Low-Dimensional Manifolds (Proc. Second Sussex Conf., Chelwood Gate, 1977), Lecture Notes in Math., Vol. 722, Springer, Berlin (1979), 48–66.

[65] T. Jørgensen, *On 3-manifolds fibering over the circle*, Ann. of Math. **106** (1977), 61–72.

[66] J. Kalliongis and D. McCullough, *π_1-injective mappings of compact 3-manifolds*, Proc. London Math. Soc. **52** (1986), 173–192.

[67] J. Kalliongis and D. McCullough, *Maps inducing isomorphisms on fundamental groups of compact 3-manifolds*, J. London Math. Soc. **35** (1987), 177–192.

[68] J. Kalliongis and D. McCullough, *Homeotopy groups of irreducible 3-manifolds which may contain two-sided projective planes*, Pacific J. Math. **153** (1992), 85–117.

[69] J. Kalliongis and D. McCullough, *Isotopies of 3-manifolds*, Topology Appl. **71** (1996), 227–263.

[70] M. Kapovich, *Hyperbolic Manifolds and Discrete Groups*, Progress in Math., Vol. 183, Birkhäuser (2001).

[71] H. Kneser, *Geschlossene Flächen in dreidimensionalen Mannigfaltigkeiten*, Jahr. Deutsch. Math. Verein. **38** (1929), 248–260.

[72] S. Kojima, *A construction of geometric structures on Seifert fibered spaces*, J. Math. Soc. Japan **36** (1984), 483–495.

[73] F. Laudenbach, *Topologie de la Dimension* 3; *Homotopie et Isotopie*, Astérisque, Vol. 12, Société Mathématique de France (1974).
[74] W.B.R. Lickorish, *A representation of orientable combinatorial* 3-*manifolds*, Ann. of Math. **76** (1962), 531–540.
[75] R. Kulkarni, K.B. Lee and F. Raymond, *Deformation spaces for Seifert manifolds*, Geometry and Topology, Maryland 1983, Lecture Notes in Math., Vol. 1167, J. Alexander and J. Harer, eds, Springer, Berlin (1983), 235–266.
[76] C. McMullen, *Amenability, Poincaré series and holomorphic averaging*, Invent. Math. **97** (1989), 95–127.
[77] C. McMullen, *Iteration on Teichmüller space*, Invent. Math. **99** (1990), 425–454.
[78] C. McMullen, *Renormalization and* 3-*Manifolds which Fiber over the Circle*, Annals of Math. Studies, Vol. 142, Princeton University Press (1994).
[79] C. McMullen, *Complex Dynamics and Renormalization*, Annals of Math. Studies, Vol. 135, Princeton University Press (1996).
[80] W. Magnus, *Residually finite groups*, Bull. Amer. Math. Soc. **75** (1965), 305–316.
[81] W. Magnus, A. Karrass and D. Solitar, *Combinatorial Group Theory*, Wiley (1966).
[82] A.I. Mal'cev, *On isomorphic matrix representations of infinite groups*, Mat. Sb. **8** (1940), 405–422 (Russian); English translation: Amer. Math. Soc. Transl. **45** (1965), 1–18.
[83] B. Maskit, *Kleinian Groups*, Grundlehren der Math. Wissenschaften, B. 287, Springer, Berlin (1988).
[84] W.S. Massey, *A Basic Course in Algebraic Topology*, Springer, Berlin (1991).
[85] R.C. McOwen, *Point singularities and conformal metrics on Riemann surfaces*, Proc. Amer. Math. Soc. **103** (1988), 222–224.
[86] Y. Minsky, *On rigidity, limit sets and end invariants of hyperbolic* 3-*manifolds*, J. Amer. Math. Soc. **7** (1994), 539–588.
[87] W.H. Meeks and G.P. Scott, *Finite group actions on* 3-*manifolds*, Invent. Math. **86** (1986), 287–346.
[88] W.H. Meeks, L. Simon and S.-T. Yau, *Embedded minimal surfaces, exotic spheres, and manifolds with positive Ricci curvature*, Ann. of Math. **116** (1982), 621–659.
[89] W.H. Meeks and S.-T. Yau, *Topology of three-dimensional manifolds and the embedding problems in minimal surface theory*, Ann. of Math. **112** (1980), 441–485.
[90] W.H. Meeks and S.-T. Yau, *The classical Plateau problem and the topology of three-dimensional manifolds*, Topology **21** (1982), 409–442.
[91] W.H. Meeks and S.T. Yau, *Group actions on* $\mathbb{R}^3$, The Smith Conjecture, H. Bass and J.W. Morgan, eds, Academic Press, New York (1984), 167–179.
[92] G. Mess, *Centers of* 3-*manifold groups which are coarse quasi-isometric to planes*, Unpublished preprint, U.C.L.A. (1988).
[93] R. Meyerhoff, *Density of the Chern–Simons invariant for hyperbolic* 3-*manifolds*, Low-Dimensional Topology and Kleinian Groups (Coventry/Durham, 1984), London Math. Soc. Lecture Note Ser., Vol. 112, Cambridge Univ. Press (1986), 217–239.
[94] R. Meyerhoff and W.D. Neumann, *An asymptotic formula for the eta invariants of hyperbolic* 3-*manifolds*, Comment. Math. Helv. **67** (1992), 28–46.
[95] J. Milnor, *A unique factorization for* 3-*manifolds*, Amer. J. Math. **84** (1962), 1–7.
[96] J.-M. Montesinos, *Variedades de Seifert que son recubridores cíclicos ramificados de dos hojas*, Bol. Soc. Mat. Mex **18** (1973), 1–32.
[97] J.-M. Montesinos, *Classical Tessellations and Three-Manifolds*, Universitext Series, Springer, Berlin (1987).
[98] J.W. Morgan, *On Thurston's uniformization theorem for the three-dimensional manifolds*, The Smith Conjecture, H. Bass and J.W. Morgan, eds, Academic Press, New York (1984), 37–125.
[99] G. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Ann. Math. Studies, Vol. 78, Princeton University Press (1973).
[100] S. Negami, *Irreducible* 3-*manifolds with non-trivial* π_2, Yokohama Math. J. **29** (1981), 133–144.
[101] W.D. Neumann and F. Raymond, *Seifert manifolds, plumbing,* μ*-invariant and orientation-reversing maps*, Algebraic and Geometric Topology (Proc. Sympos. Univ. California, Santa Barbara, 1977), Lecture Notes in Math., Vol. 664, Springer, Berlin (1978), 162–195.
[102] K. Ohshika, *Teichmüller spaces of Seifert fibered manifolds with infinite* π_1, Topology Appl. **27** (1987), 75–93.

[103] W. Opechovski, *Crystallographic and Metacrystallographic Groups*, North-Holland, Amsterdam (1986).

[104] P. Orlik, *Seifert Manifolds*, Lecture Notes in Math., Vol. 291, Springer, Berlin (1972).

[105] P. Orlik, E. Vogt and H. Zieschang, *Zur Topologie gefaserter dreidimensionaler Mannigfaltigkeiten*, Topology **6** (1967), 49–64.

[106] J.-P. Otal, *Le Théorème d'Hyperbolisation pour les Variétés Fibrées de Dimension* 3, Astérisque, Vol. 235, Société Mathématique de France (1996).

[107] J.-P. Otal, *Thurston's hyperbolization of Haken manifolds*, Surveys in Differential Geometry, Vol. III (Cambridge, MA, 1996), International Press (1998), 77–194.

[108] M. Ouyang, *A simplicial formula for the η-invariant of hyperbolic* 3*-manifolds*, Topology **36** (1997), 411–421.

[109] C.D. Papakyriakopoulos, *On Dehn's lemma and the asphericity of knots*, Ann. of Math. **66** (1957), 1–26.

[110] E. Reyssat, *Quelques Aspects des Surfaces de Riemann*, Birkhäuser (1989).

[111] R. Riley, *Parabolic representations of knot groups*, Proc. London Math. Soc. **24** (1972), 217–242.

[112] R. Riley, *A quadratic parabolic group*, Math. Proc. Cambridge Phil. Soc. **77** (1975), 181–188.

[113] R. Riley, *An elliptical path from parabolic representations to hyperbolic structures*, Topology of Low Dimensional Manifolds, Lecture Notes in Math., Vol. 722, R. Fenn, ed., Springer, Berlin (1979), 99–133.

[114] R. Riley, *Seven excellent knots*, Low-dimensional Topology, London Math. Soc. Lecture Notes Series, Vol. 48, R. Brown and T. Thickstun, eds, Cambridge University Press (1982), 81–151.

[115] R. Riley, *Applications of a computer implementation of Poincaré's theorem on fundamental polyhedra*, Math. Comp. **40** (1983), 607–632.

[116] D. Rolfsen, *Knots and Links*, Publish or Perish Inc. (1976).

[117] J.H. Rubinstein, *Free actions of some finite groups on S^3*, Math. Ann. **240** (1979), 165–175.

[118] J.H. Rubinstein, *On* 3*-manifolds that have finite fundamental group and contain Klein bottles*, Trans. Amer. Math. Soc. **251** (1979), 129–137.

[119] J.H. Rubinstein, *Nonorientable surfaces in some non-Haken* 3*-manifolds*, Trans. Amer. Math. Soc. **270** (1982), 503–524.

[120] J.H. Rubinstein and J.S. Birman, *One-sided Heegaard splittings and homeotopy groups of some* 3*-manifolds*, Proc. London Math. Soc. **49** (1984), 517–536.

[121] I. Satake, *On a generalization of the notion of manifold*, Proc. Nat. Acad. Sci. USA **42** (1956), 359–363.

[122] I. Satake, *The Gauss–Bonnet theorem for* V*-manifolds*, J. Math. Soc. Japan **9** (1957), 464–492.

[123] H. Schubert, *Knoten und Vollringen*, Acta Math. **90** (1953), 131–286.

[124] P. Scott, *A new proof of the annulus and torus theorems*, Amer. J. Math. **102** (1980), 241–277.

[125] P. Scott, *There are no fake Seifert spaces with infinite π_1*, Ann. of Math. **117** (1983), 35–70.

[126] P. Scott, *The geometries of* 3*-manifolds*, Bull. London Math. Soc. **15** (1983), 401–487.

[127] P. Scott, *Homotopy implies isotopy for some Seifert fibre spaces*, Topology **24** (1985), 341–351.

[128] H. Seifert, *Topologie dreidimensionaler gefaserter Räume*, Acta Math. **60** (1933), 147–238.

[129] H. Seifert and W. Threlfall, *Lehrbuch der Topologie*, Leipzig (1934).

[130] L.C. Siebenmann, *A do-it-yourself crystal classification*, Unpublished preprint, Université de Paris-Sud (1983).

[131] I.M. Singer, *Infinitesimally homogeneous spaces*, Comm. Pure Appl. Math. **13** (1960), 685–697.

[132] S. Smale, *Diffeomorphisms of the* 2*-sphere*, Proc. Amer. Math. Soc. **10** (1959), 621–626.

[133] P.A. Smith, *A theorem for fixed points of periodic transformations*, Ann. of Math. **35** (1934), 572–578.

[134] T. Soma, K. Ohshika and S. Kojima, *Towards a proof of Thurston's geometrization theorem for orbifolds*, Hyperbolic Geometry and 3-Manifolds (Proc. Symp. Research Inst. for Math. Sciences, Kyoto, 1985), Sûrikaisekikenkyûsho Kôkyûroku **568** (1985), 1–72.

[135] D.P. Sullivan, *Travaux de Thurston sur les groupes quasi-fuchsiens et sur les variétés hyperboliques de dimension* 3 *fibrées sur S^1*, Séminaire Bourbaki 554, Lecture Notes in Math., Vol. 842, Springer, Berlin (1981), 196–214.

[136] Y. Takeuchi, *Waldhausen's classification theorem for finitely uniformizable* 3*-orbifolds*, Trans. Amer. Math. Soc. **328** (1991), 151–200.

[137] W. Threlfall and H. Seifert, *Topologische Untersuchung der diskontinuitätbereiche endlichen Bewegungsgruppen des dreidimensionalen sphärischen Raumes* I, Math. Ann. **104** (1930), 1–30; II, Math. Ann. **107** (1932), 543–586.

[138] W.P. Thurston, *The Topology and Geometry of 3-Manifolds*, Lecture Notes, Princeton University (1976–79).

[139] W.P. Thurston, *Hyperbolic geometry and 3-manifolds* (Notes by P. Scott), Low-Dimensional Topology, London Math. Soc. Lecture Notes Series, Vol. 48, R. Brown and T. Thickstun, eds, Cambridge University Press (1982), 9–25.

[140] W.P. Thurston, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. **6** (1982), 357–381.

[141] W.P. Thurston, *Hyperbolic structures on 3-manifolds* I. *Deformation of acylindrical manifolds*, Ann. of Math. **124** (1986), 203–246.

[142] W.P. Thurston, *Hyperbolic structures on 3-manifolds* II. *Surface groups and manifolds which fiber over the circle*, Unpublished preprint, Princeton University (1986).

[143] W.P. Thurston, *The Topology and Geometry of 3-Manifolds*, S. Levy, ed., Princeton University Press (1997).

[144] F. Waldhausen, *Eine Klasse von dreidimensionalen Mannigfaltigkeiten* I, Invent. Math. **3** (1967), 308–333; II, ibid. **4** (1967), 87–117.

[145] F. Waldhausen, *On irreducible 3-manifolds which are sufficiently large*, Ann. of Math. **87** (1968), 56–88.

[146] F. Waldhausen, *Über Involutionen der 3-Sphäre*, Topology **8** (1969), 81–91.

[147] F. Waldhausen, *On the determination of some bounded 3-manifolds by their fundamental group alone*, Proc. of the International Symposium on Topology and its Applications, Herceg-Novi 1968, Beograd (1969), 331–332.

[148] F. Waldhausen, *Recent results on sufficiently large 3-manifolds*, Proceedings of Symposia in Pure Mathematics, Vol. 32, Part 2, R.J. Milgram, ed., Amer. Math. Soc., Providence, RI (1978), 21–38.

[149] C.T.C. Wall, *Geometries and geometric structures in real dimension 4 and complex dimension 2*, Geometry and Topology (College Park, Md., 1983/84), Lecture Notes in Math., Vol. 1167, Springer, Berlin (1985), 268–292.

[150] C.T.C. Wall, *Geometric structures on compact complex analytic surfaces*, Topology **25** (1986), 119–153.

[151] A.H. Wallace, *Modifications and cobounding manifolds*, Can. J. Math. **12** (1960), 503–528.

[152] J. Weeks, *Convex hulls and isometries of cusped hyperbolic 3-manifolds*, Topology Appl. **52** (1993), 127–149.

[153] J. Weeks et al., *SnapPea*, Computer program, Geometry Center, University of Minnesota (1990–97).

[154] J.A. Wolf, *Spaces of Constant Curvature*, McGraw-Hill (1967).

[155] T. Yoshida, *The η-invariant of hyperbolic 3-manifolds*, Invent. Math. **81** (1985), 473–514.

[156] Q. Zhou, *3-dimensional geometric cone structures*, Doctoral thesis, U.C.L.A. (1989).

[157] Q. Zhou, *The moduli space of hyperbolic cone structures*, J. Differential Geom. **51** (1999), 517–550.

CHAPTER 4

Dehn Surgery on Knots

Steven Boyer*

Département de mathématiques, UQAM, P.O. Box 8888, Station Centre-ville, Montréal H3C 3P8, Québec, Canada

Contents

*Partially supported by grants: NSERC OGP 0009446 and FCAR ER-68657.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

How far is a closed, connected, oriented 3-manifold W from being homeomorphic to S^3? To explain one possible response to this question we consider a Heegaard splitting of W. This is a homeomorphism $W \cong H_1 \cup_f H_2$ where H_1, H_2 are two oriented solid handlebodies of genus g and $f : \partial H_1 \to \partial H_2$ is an orientation reversing homeomorphism (for a detailed discussion of Heegaard splittings of compact 3-manifolds see Chapter 18). The 3-sphere also admits a genus g Heegaard splitting, say $S^3 = H_1 \cup_{f_0} H_2$.

Lickorish [68] proved that any orientation preserving homeomorphism of the boundary of a solid handlebody extends to a homeomorphism of the complement of a finite collection of disjoint open solid tori in the interior of the handlebody. In particular there are disjoint solid tori $V_1, V_2, \ldots, V_m \subset \text{int}(H_2)$ such that ff_0^{-1} extends to a homeomorphism $h : H_2 \setminus \text{int}(V_1 \cup V_2 \cup \cdots \cup V_m) \to H_2 \setminus \text{int}(V_1 \cup V_2 \cup \cdots \cup V_m)$. It follows that h pieces together with the identity of H_1 to provide a homeomorphism between $H_1 \cup_{f_0} (H_2 \setminus \text{int}(V_1 \cup V_2 \cup \cdots \cup V_m))$ and $H_1 \cup_f (H_2 \setminus \text{int}(V_1 \cup V_2 \cup \cdots \cup V_m))$, and thus the exterior of some link L in the 3-sphere is homeomorphic to the exterior of a link in W. Put another way, W may be obtained by removing a tubular neighbourhood of the link L from S^3 and gluing it back in a possibly different fashion. The link $L \subset S^3$ then measures the obstruction to extending the identity of H_1 to a homeomorphism from S^3 to W.

This operation of removing a solid torus from a 3-manifold and then reattaching it in some fashion, known as a *Dehn surgery*, has become one of the fundamental methods used to represent 3-dimensional manifolds. It was introduced by Dehn [25] as a method for constructing "Poincaré spaces", that is non-simply-connected 3-manifolds which possess the same homology as the 3-sphere. As we have just observed, any closed, connected, orientable 3-manifold may be obtained from the 3-sphere by Dehn surgery on a link $L \subset S^3$, or equivalently, by a finite sequence of surgeries on knots in 3-manifolds. Thus as a means to study 3-manifold topology, it is of interest to determine how the topological and geometric properties of a knot K lying in the interior of an orientable 3-manifold W are related to the analogous properties of the surgered manifold.

Let $N(K) \subset \text{int}(W)$ be a closed tubular neighbourhood of K and consider the manifold $M_K = W \setminus \text{int}(N(K))$, referred to as the *exterior* of K. Many of the results and examples relating to Dehn surgery which have appeared over the last two decades show that the topological and geometric properties of the spaces M_K persist in the manifolds resulting from surgery on K, at least for generic surgeries. That exceptional surgeries arise often is easy to see – a knot in the 3-sphere with an arbitrarily complicated exterior always has a "trivial" surgery which yields S^3. What is rare is for a fixed knot to admit more than a few such surgeries, and the main thrust of research has been to describe these exceptional phenomena. We take this work as our central theme.

Though most of the major methods of low-dimensional topology have been brought to bear on surgery theory, we shall focus on those which have yielded the most profound results relating to our theme. These include

(i) the geometric methods of Thurston et al. [95–97,99];

(ii) the combinatorial/topological theory of intersection graphs largely developed by C.McA. Gordon and J. Luecke. A survey of this subject may be found in [44];

(iii) the character variety methods pioneered by Casson [4] and Culler and Shalen ([92], Chapter 19 of this volume);
(iv) the theory of 2-dimensional foliations in 3-manifolds and sutured manifold theory as developed by Gabai [32–34] and Scharlemann [89], and the theory of essential laminations in 3-manifolds introduced by Gabai and Oertel in [38]. An overview of this subject can be found in [37].

We shall restrict ourselves, for the most part, to discussing the topology of the Dehn surgery operation, directing the reader to Chapters 3 and 17 for a description of some of its geometric aspects.

We greatly profited from the earlier survey articles of Gordon [42,43], Luecke [71], and Gabai [37] which touched on various aspects of our topic, and recommend them to the reader. R. Kirby's problem list [64] also contains much material and discussion of interest. D. Rolfsen's text [85] remains an excellent introduction to knot theory and surgery.

Throughout this chapter we shall work in both the smooth category and the equivalent tame-topological category, and assume that our 3-manifolds are orientable. Knots and links will be assumed to lie in the interiors of compact, connected, orientable 3-manifolds. Surfaces in 3-manifolds will be assumed to be either properly embedded, or subsets of the boundary. Regular neighbourhoods will be be denoted by $N(\cdot)$. In particular, a closed tubular neighbourhood of a link L will be denoted by $N(L)$. The *exterior* of $L \subset W$ is the compact manifold $M_L = W \setminus \mathrm{int}(N(L))$. It depends only on L, at least up to an orientation preserving homeomorphism.

We have organized this chapter as follows. In Section 2 we introduce slopes, the basic parameter of the surgery operation, and describe several distinguished families of slopes. In Section 3 we discuss the existence and uniqueness of surgery presentations of a given 3-manifold. The final three sections form the heart of the chapter and focus on exceptional surgeries. Thus we examine the effect of surgery on various sorts of surfaces in Section 4, the relationship between the global properties of a knot's exterior and the manifolds obtained by surgery on the knot in Section 5, and we specialize to exceptional surgery on knots in the 3-sphere in Section 6.

There are many uncited questions and conjectures which appear below. These are, to the best of my knowledge, most appropriately attributed to folklore.

The author would like to thank Cameron McA. Gordon and Daniel Lines who commented on an earlier version of this paper, and Corinne and Paul Libbrecht who prepared the figures which appear here.

2. Slopes

The surgery operation is a two-stage construction. First a tubular neighbourhood of a knot is removed from the ambient 3-manifold (*drilling*), and then it is reattached (*filling*). Let us consider the second stage more closely.

Take T to be a toral boundary component of an orientable 3-manifold M. Given any homeomorphism $f : \partial(S^1 \times D^2) \to T$, form the identification space $M(T; f) = (S^1 \times D^2) \cup_f M$ obtained by identifying the points of $\partial(S^1 \times D^2)$ with their images by f. We refer to $M(T; f)$ as a *(Dehn) filling* of M along T. A Dehn surgery on a knot K is then a filling of M_K along $\partial N(K)$.

In this section we introduce the notion of a slope on a torus, the basic parameter of the filling operation.

2.1. *Fillings and slopes*

Fix a 3-manifold M and a torus $T \subset \partial M$. By the nature of its construction, a filling $M(T; f)$ depends only on the isotopy class of the attaching homeomorphism $f : \partial(S^1 \times D^2) \to T$. In fact the dependence on f is much weaker, for if $C_0 = \{pt\} \times \partial D^2 \subset \partial(S^1 \times D^2)$, then $M(T; f)$ depends only on the isotopy class of the curve $f(C_0)$ in T. To see this, let $D_0 = \{pt\} \times D^2 \subset S^1 \times D^2$ and observe that $S^1 \times D^2$ splits into two pieces A and B, where A is a closed tubular neighbourhood of D_0 and B is the 3-ball $\overline{S^1 \times D^2 \setminus A}$. Now think of $M(T; f)$ as being built in two stages. In the first we form $A \cup M$, which amounts to attaching a 2-handle to M along a tubular neighbourhood of $f(C_0) \subset T$. Now a manifold obtained from such an attachment depends only on the isotopy class in T of the attaching 1-sphere, which in our case is $f(C_0)$. In the second stage we form $M(T; f) = B \cup (A \cup M)$ by attaching the 3-ball B to $A \cup M$ along its 2-sphere boundary. As any homeomorphism of a 2-sphere extends over the 3-ball (e.g., by coning), $M(T; f)$ is completely determined by $A \cup M$, and hence by the isotopy class of $f(C_0)$ in T. Figure 1 depicts a possible $f(C_0)$ in the case where M is the exterior of the right-handed trefoil knot.

This analysis indicates the importance of understanding the isotopy classes of simple closed curves on a torus. In fact these classes are particularly well-behaved, as the following lemma indicates. Proofs of its assertions may be found in Chapter 2.C of Rolfsen's text [85].

LEMMA 2.1. *Let T be a torus.*

(i) *A separating simple closed curve on T bounds a 2-disk in T.*

(ii) *For any essential, simple closed curve C on T, there is a dual simple closed curve C' which intersects C exactly once and transversely.*

(iii) *Disjoint essential, simple closed curves on T are parallel, that is they cobound an annulus embedded in T.*

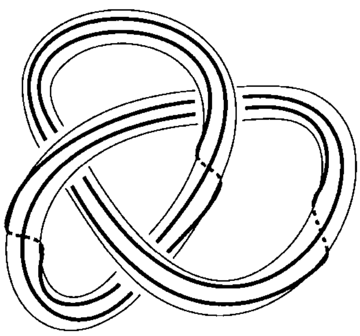


Fig. 1. A possible $f(C_0)$.

(iv) *Two oriented essential simple closed curves on T are isotopic if and only if the 1-cycles they define are homologous.*
(v) *A homology class in $H_1(T)$ is represented by the fundamental class of an oriented, essential, simple closed curve if and only if it is primitive, that is if and only if it is an element of some basis of $H_1(T)$.*
(vi) *Given any two essential, simple closed curves C, C' on T, there is an orientation preserving homeomorphism $f : T \to T$ such that $f(C) = C'$.*

DEFINITION 2.2. A *slope* on a torus T is the isotopy class of an essential, unoriented, simple closed curve on T. The set of slopes on T will be denoted by *Slope*(T). Two slopes r_1, r_2 on T are called *dual* if they have representative curves which intersect exactly once and transversely. Finally if K is a knot in a 3-manifold W, then a *slope of K* is any slope on $\partial N(K)$.

We summarize the discussion above in the next proposition.

PROPOSITION 2.3. *A Dehn filling of M along a torus $T \subseteq \partial M$ is determined, up to orientation preserving homeomorphism, by a slope on T. Furthermore, any slope on T arises as the slope of a Dehn filling of M.*

Denote by $M(T; r)$ any Dehn filling of M along T corresponding to a gluing homeomorphism f for which $f(C_0)$ represents the slope r. When $T = \partial M$ we shall abbreviate this notation to $M(r)$.

There is one distinguished slope determined by any knot. A *meridian* for a knot $K \subset W$ is any essential, simple closed curve on $\partial N(K)$ which is homologically trivial in $N(K)$ (such curves actually bound 2-disks in $N(K)$). Meridians are well-defined up to isotopy (cf. Lemma 2.1) and so determine a slope μ_K of K, called the *meridional* slope of K.

The *trivial* Dehn surgery on a knot $K \subset W$ is the surgery corresponding to the meridional slope. Evidently $M_K(\mu_K) \cong W$, for in this case we can equate $S^1 \times D^2$ with $N(K)$ and then choose the gluing map so that its effect in attaching $S^1 \times D^2$ to M_K is just to return $N(K)$ to W.

There is another distinguished slope for knots in the 3-sphere, or more generally for null-homologous knots in an arbitrary orientable 3-manifold W. A null-homology of a knot $K \subset W$ can be realized by a compact, connected, orientable, smooth subsurface of W whose boundary is K. Such a surface, called a *Seifert surface* of K, may be isotoped to intersect $N(K)$ in an annulus whose boundary consists of K and an essential, simple closed curve lying on $\partial N(K)$. The latter curve, called a *longitude* of K, is characterized up to isotopy on $\partial N(K)$ by the fact that it is essential on $\partial N(K)$ while homologically trivial in the exterior of K (cf. Lemma 2.1). The *longitudinal* slope of K, denoted by λ_K, is the slope of any longitude of K. It is evident that μ_K and λ_K are dual slopes (cf. Definition 2.2).

2.2. *Parameterizing slopes*

It follows from Lemma 2.1(iv), (v) that the set of slopes on a torus T corresponds bijectively to the set of $\pm$ pairs of primitive classes in $H_1(T)$. Explicitly, if we choose any representative for a slope r and orient it, the fundamental homology class of this oriented circle determines a primitive homology class $\alpha \in H_1(T)$. Changing orientation changes the sign of α. We shall call $\pm\alpha$ the homology classes *carried by* r. These observations lead to a parameterization of the set of slopes.

PROPOSITION 2.4. *Let T be a torus and set $P^1(\mathbf{Q}) = \mathbf{Q} \cup \{\frac{1}{0}\}$. Each choice of ordered basis for $H_1(T)$ determines a bijection between the set of slopes Slope(T) on T and $P^1(\mathbf{Q})$. If K is a null-homologous knot in the interior of an oriented 3-manifold and $T = \partial N(K)$, then the correspondence can be made canonical.*

PROOF. Fix an ordered basis $\{\alpha, \beta\}$ for $H_1(T)$. Any slope r on T determines a pair of relatively prime integers p, q such that the homology classes in $H_1(T)$ carried by r are $\pm(p\alpha + q\beta)$. This correspondence gives rise to the desired bijection *Slope*$(T) \leftrightarrow P^1(\mathbf{Q})$ via $r \leftrightarrow \pm(p\alpha + q\beta) \leftrightarrow p/q$.

Suppose now that K is a null-homologous knot in the interior of an oriented 3-manifold W and note that the induced orientation on $N(K) \subset W$ determines an orientation on T. Choose classes $\alpha(\mu_K), \alpha(\lambda_K) \in H_1(T)$ carried by μ_K, λ_K. Since μ_K and λ_K are dual slopes, the algebraic intersection $\alpha(\mu_K) \cdot \alpha(\lambda_K)$ is either $+1$ or -1. If we require that this intersection be $+1$, then up to a simultaneous change of sign, $\{\alpha(\mu_K), \alpha(\lambda_K)\}$ is a well-defined ordered basis of $H_1(T)$. Thus the correspondence $r \leftrightarrow \pm(p\alpha(\mu_K) + q\alpha(\lambda_K)) \leftrightarrow p/q$ becomes canonical. □

We shall always assume that S^3 is given its usual orientation based on the right-hand rule. Thus slopes of knots in the 3-sphere are canonically identified with $P^1(\mathbf{Q})$.

DEFINITION 2.5. Let K be a knot in S^3. For a slope r of K corresponding to the fraction $p/q \in P^1(\mathbf{Q})$, $M_K(r)$ will also be denoted by $M_K(p/q)$. An *integral slope* of K is a slope corresponding to an integer, and an *integral* surgery on K is a surgery whose slope is integral.

Figure 1 depicts a representative curve for the 9/2 slope of the right-handed trefoil knot.

EXAMPLE 2.6. Let $K \subset S^3$ be the trivial knot. We determine the manifolds $M_K(p/q)$, $p \geqslant 0$. There is a canonical identification $M_K \equiv S^1 \times D^2$ for which $S^1 \times \{1\} \subset \partial M_K$ represents μ_K and $\{1\} \times \partial D^2 \subset \partial M_K$ represents λ_K. Hence $M_K(p/q)$ is the union $S^1 \times D^2 \cup_f S^1 \times D^2$ where the meridian $\{1\} \times \partial D^2$ of the first $S^1 \times D^2$ is identified via f with a curve on the boundary of the second $S^1 \times D^2$ which is homologous to the sum of p copies of $S^1 \times \{1\}$ and q copies of $\{1\} \times \partial D^2$. Thus if $L(p, q)$ denotes the (p, q) lens space (see, e.g., [85, p. 233]),

$$M_K(p/q) \cong \begin{cases} S^1 \times S^2 & \text{if } p = 0, \\ S^3 & \text{if } p = 1, \\ L(p, q) & \text{otherwise.} \end{cases}$$

Other worked examples may be found in [85, Chapter 9].

2.3. *The surgery plane and the distance between slopes*

For many purposes it is useful to think of the set of slopes on a torus T as $\pm$ pairs of primitive elements of the lattice $H_1(T)$ lying in the *surgery plane* $H_1(T; \mathbf{R})$. If an ordered basis for $H_1(T)$ is chosen, we may identify $H_1(T) \equiv \mathbf{Z}^2 \subset H_1(T; \mathbf{R}) \equiv \mathbf{R}^2$. The slopes on T then correspond to the pairs $\pm(p, q)$ of primitive lattice points, and therefore to the set of lines through the origin in $\mathbf{R}^2$ having a rational slope (hence the origin of the term *slope*).

Many of the results on Dehn surgery are most naturally couched in terms of the following notion.

DEFINITION 2.7. The *distance* between two slopes r_1 and r_2 on a torus T is their *geometric intersection number*

$$\Delta(r_1, r_2) = \min\{\#(C_1 \cap C_2) \mid C_i \text{ is a simple curve which represents } r_i,\ i = 1, 2\}.$$

Basic surface topology shows that

$$\Delta(r_1, r_2) = |\alpha(r_1) \cdot \alpha(r_2)|,$$

the absolute value of the algebraic intersection of primitive homology classes carried by r_1 and r_2. Thus if we fix a basis $\{\alpha, \beta\}$ for $H_1(T)$ and write $\alpha(r_i) = \pm(p_i\alpha + q_i\beta)$, then

$$\Delta(r_1, r_2) = |p_1q_2 - p_2q_1|.$$

Though not a distance in the sense of metric spaces, $\Delta(\cdot, \cdot)$ nevertheless has many useful properties. For instance $\Delta(r_1, r_2) = 0$ if and only if $r_1 = r_2$ and $\Delta(r_1, r_2) = 1$ if and only if $\{\alpha(r_1), \alpha(r_2)\}$ is a basis for $H_1(T)$.

The geometry of $\Delta(\cdot, \cdot)$ becomes clearer when we consider it as the restriction of the function

$$\Delta : H_1(T; \mathbf{R}) \oplus H_1(T; \mathbf{R}) \to [0, \infty)$$

given by $\Delta(v, w) = |v \cdot w|$. For a fixed $v_0 \neq 0$ in the surgery plane and $t > 0$, $\{v \in H_1(T; \mathbf{R}) \mid \Delta(v, v_0) \leqslant t\}$ is an infinite band parallel to v_0. It follows that if the distances between the elements of a set of slopes is bounded by some positive integer n, then the set of homology classes associated to this set lie in a certain bounded region of the surgery plane. More precisely we have the following lemma, whose elementary proof may be found in [46].

LEMMA 2.8. *Let n be a non-negative integer and let $\mathcal{S}_0$ be a family of slopes on a torus T which satisfies $\Delta(r_1, r_2) \leqslant n$ for each pair $r_1, r_2 \in \mathcal{S}_0$. Then there is a choice $\{\alpha, \beta\}$ of ordered basis for $H_1(T)$ such that $\mathcal{S}_0$ corresponds to a subset of the collection of primitive lattice points in $\{\pm(1, 0)\} \cup \{\pm(p, q) \mid 0 \leqslant p \leqslant q \leqslant n\}$.*

Using this lemma it can be verified that for $0 \leqslant n \leqslant 10$, the maximal number of slopes in a set of slopes $\mathcal{S}_0$, as in Lemma 2.8, is given by Table 1.

Table 1

n	0	1	2	3	4	5	6	7	8	9	10
Upper bound for $\#\mathcal{S}_0$	1	3	4	6	6	8	8	10	12	12	12

2.4. *Essential surfaces and boundary slopes*

In this section we discuss an important class of slopes on a toral boundary component T of a compact orientable 3-manifold M, whose properties are closely related to the topology of the fillings $M(T; r)$.

Consider an orientable surface F in M, not homeomorphic to a disk or a sphere, which is either properly embedded or a subset of ∂M. We shall say that F is *incompressible* if for each 2-disk D embedded in M such that $D \cap F = \partial D$, there is a disk D' embedded in F such that $\partial D' = \partial D$. The loop theorem implies that F is incompressible if and only if the natural homomorphism $\pi_1(F) \to \pi_1(M)$ is injective for each choice of basepoint in F.

A 3-manifold M is called *irreducible* if each 2-sphere embedded in M bounds a 3-ball, and *reducible* otherwise. It is called ∂*-irreducible* if ∂M is incompressible in M, and ∂*-reducible* otherwise.

There is a relative form of incompressibility which often arises. A properly embedded, orientable surface F in M is called ∂*-incompressible* if for each disk $D \subset M$ satisfying

- $\partial D = \alpha \cup \beta$ where α and β are arcs meeting only in their endpoints,
- $\alpha = D \cap F$ and $\beta = D \cap \partial M$,

there is a disk $D' \subset F$ with $\alpha \subset \partial D'$ and $\partial D' \setminus \alpha \subset \partial F$. This notion can be characterized as follows: F is ∂-incompressible if and only if the function $\pi_1(F, \partial F) \to \pi_1(M, \partial M)$ is injective for each choice of basepoint.

There are general criteria which guarantee that certain incompressible surfaces are also ∂-incompressible. Straightforward arguments shows that if M is irreducible and F is a properly embedded, incompressible surface whose boundary lies in a toral component of ∂M, then F is either ∂-incompressible or a boundary-parallel annulus.

By an *essential* surface in a compact, orientable 3-manifold M we mean a properly embedded orientable surface F in M satisfying

- no component of F is parallel into ∂M,
- no 2-sphere component of F bounds a 3-ball in M,
- no 2-disk component of F cobounds a 3-ball in M with a disk in ∂M,
- each non-simply-connected component of F is both incompressible and ∂-incompressible.

A slope r on a toral boundary component T of a 3-manifold M is called a *boundary slope* if there is an essential surface $F \subset M$ whose boundary lies on T and consists of a non-empty set of parallel simple closed curves of slope r.

The fundamental result concerning these slopes is contained in the following theorem of A. Hatcher.

THEOREM 2.9 (Hatcher [54]). *Let M be a compact, orientable, irreducible, ∂-irreducible 3-manifold and T a toral boundary component of ∂M. Then there are only finitely many boundary slopes on T.*

As an example, recall (Section 2.1) the longitudinal slope λ_K of a knot K in the 3-sphere. The loop theorem implies that a Seifert surface for K may be made essential by a sequence of compressions, and so λ_K is a boundary slope of M_K. It is a hard theorem of Culler and Shalen [22] that if K is knotted, then there are other boundary slopes on ∂M_K.

The problem of determining the set of boundary slopes on $T \subseteq \partial M$ is quite difficult in general, though several interesting families have been studied. In particular Hatcher and Thurston [56] determined those for 2-bridge knots, while Hatcher and Oertel [55] produced an algorithm for studying the boundary slopes of Montesinos knots.

3. Realizing 3-manifolds by Dehn surgery

The questions we examine in this section concern the existence and uniqueness of representations of a given closed, connected, orientable 3-manifold by surgery on a finite number of knots in the 3-sphere.

3.1. *Existence. The Lickorish–Wallace theorem*

By a set of *surgery data* $(L; r_1, r_2, \ldots, r_n)$ we mean a link $L = K_1 \cup \cdots \cup K_n$ lying in the interior of a 3-manifold W together with a slope r_i for each of its components K_i. We denote by $L(r_1, r_2, \ldots, r_n)$ the manifold obtained by performing the Dehn surgeries prescribed by the surgery data. In the special case that $W = S^3$ and each r_i is an integral slope, the surgery data $(L; r_1, r_2, \ldots, r_n)$ is often referred to as a *framed link.*

The following result is known as the *fundamental theorem of surgery theory.*

THEOREM 3.1 (Lickorish [68], Wallace [104]). *Let W be a closed, connected, orientable 3-manifold. Then one may find a framed link $(L; r_1, \ldots, r_n)$ in S^3 such that W is homeomorphic to $L(r_1, r_2, \ldots, r_n)$.*

PROOF. We sketched a proof in the introduction that W is of the form $L(r_1, \ldots, r_n)$ for some link L in the 3-sphere. The point that remains to be verified is that the slopes $r_1, r_2, \ldots, r_n$ may be chosen to be integral.

Recall that the key to the argument in the introduction was that a certain orientation preserving homeomorphism f of the boundary F of a solid handlebody H extends to a homeomorphism f^* of the exterior H_0 of some link $L \subset H$. Referring to the construction, we see that the slopes of the surgery presentation of W correspond to the pre-images by f^* of the meridians of the components of L. Thus we must understand how the extension f^* is built.

Consider first of all the case where f is a Dehn twist D_C along a simple closed curve C in F. Let $C_0 \subset \text{int}(H)$ be a parallel copy of C lying just below F and let H_0 be its exterior. There is a properly embedded "vertical" annulus $A \subset H_0$ contained in a collar of F whose boundary is $C \cup C'$ where $C' \subset \partial N(C_0)$. Note that by twisting along A, D_C extends to a self-homeomorphism of H_0. The inverse of this map sends the meridional slope of C_0 to either the $+1$-slope or -1-slope, depending on the sense of the Dehn twist D_C. Hence the theorem holds in this particular situation. The general case is proven similarly, for we can invoke the *Lickorish twist theorem* [68], which states that any orientation preserving homeomorphism $g: F \to F$ is isotopic to a composition of a finite number of Dehn twists. Arguing as above, it can be shown that f extends to a homeomorphism f^* of the exterior H_0 of some link $L \subset H$ which preserves the boundary components H_0, and whose inverse sends the meridional slope of each component of L to either its $+1$-slope or -1-slope. This completes the proof. □

A different and independent approach to the proof of Theorem 3.1, due to A. Wallace, may be found in [104]. Its starting point is the fact that any closed, orientable 3-manifold W is the boundary of a smooth, compact, orientable 4-manifold (see Thom [94]). A "handle-swapping" argument is then used to show that W is the boundary of a smooth 4-manifold made up of one 0-handle and a finite number of 2-handles. As we shall see in the next section, this implies the desired result.

3.2. *Uniqueness. The rational calculus and Kirby's theorem*

A *surgery presentation* of a 3-manifold W is a set of surgery data $(L; r_1, r_2, \dots, r_n)$ where $L \subset S^3$ and $W \cong L(r_1, r_2, \dots, r_n)$. The result of the last section begs the question: to what extent is a surgery presentation of a manifold W unique? It is easy to see that the same manifold can be presented in different ways. For instance from Example 2.6 we see that if K is the trivial knot, then for each $n \in \mathbf{Z}$, $M_K(p/q) \cong L(p,q) \cong L(p, q+np) \cong M_K(p/(q+np))$. Kirby [63] introduced two moves on (integrally) framed links which do not alter the presented manifold and he succeeded in proving that two framed links present manifolds which are orientation preserving homeomorphic if and only if they are related by a finite sequence of these moves (see Theorem 3.2 below). Fenn and Rourke [30] showed that these two moves could be replaced by a single one, known as either a *Kirby move* or a *twist move*. It is depicted in Figure 2. The art of manipulating framed link diagrams according to these moves is called the *Kirby calculus*.

The problem of uniqueness of surgery presentations for general sets of surgery data was completely analysed by Rolfsen [86]. In the 1977 edition of his book [85] he had independently developed a *rational calculus* based on the following three moves. In the third, $lk(-,-)$ denotes linking number.

Isotopy. Replace $(L; p_1/q_1, \dots, p_n/q_n)$ by $(L'; p_1/q_1, \dots, p_n/q_n)$ where L' is isotopic to L by an isotopy which respects the given ordering of their component knots.

Trivial insertion or deletion. Add or remove a knot with slope $1/0$.

Twist move. Replace a set of surgery data $(L; p_0/q_0, p_1/q_1, \dots, p_n/q_n)$ where K_0 is a trivial knot by $(L'; p_0/(p_0+q_0), p_1/q_1 + (lk(K_0, K_1))^2, \dots, p_n/q_n + (lk(K_0, K_n))^2)$ as depicted in Figure 2, or vice versa.

The first two moves obviously do not change the manifold resulting from the prescribed surgery, while neither does the third. This is because by performing a right-hand twist on a meridian disk of $M_{K_0} \cong S^1 \times D^2$, one constructs a homeomorphism between the exteriors of the links L and L' which sends the chosen slopes of L to those of L'.

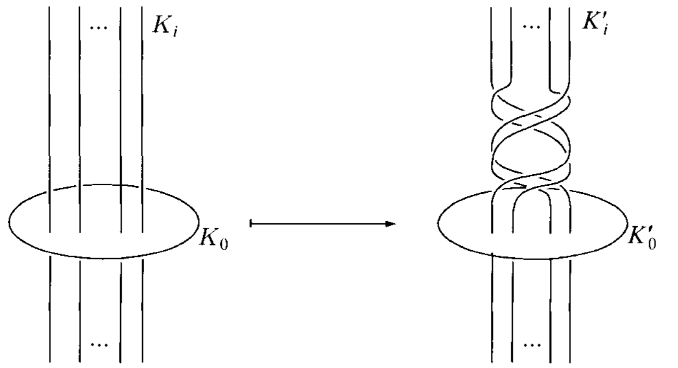


Fig. 2. The twist move.

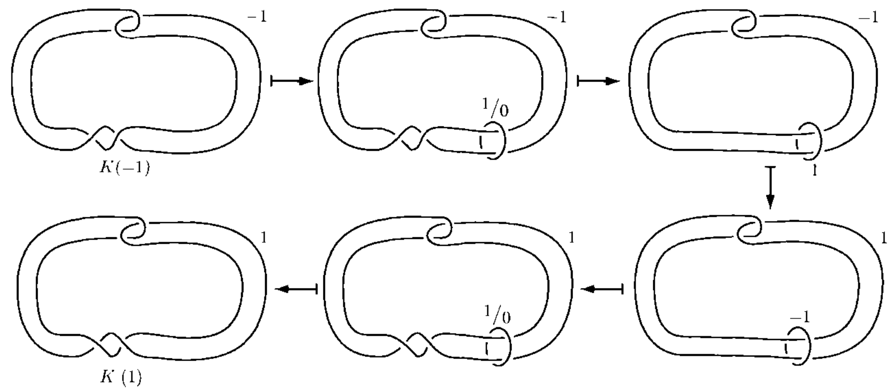


Fig. 3.

As an example let K be a right-handed trefoil knot and let K' be a figure eight knot. Figure 3 presents a pictorial proof that $M_K(-1) \cong M_{K'}(1)$. In this figure the second and sixth surgery presentations are obtained by adding or deleting a knot with slope $1/0$, the third and fifth ones are obtained by a twist move, while the fourth is obtained from the third by an isotopy.

To prove his uniqueness result, Kirby capitalized on the relationship between surgeries on framed links in the 3-sphere and the attachment of 2-handles to the 4-ball. We describe this relationship now.

When a 2-handle $(D^2 \times D^2, S^1 \times D^2)$ is attached to a 4-manifold V via an attaching map $g : S^1 \times D^2 \to \partial V$, the boundary of the new 4-manifold coincides with the surgery on ∂V along the knot $K = g(S^1 \times \{0\})$ with slope r represented by the curve $g(S^1 \times \{1\}) \subset g(S^1 \times S^1) = \partial N(K)$. Any representative curve for such a slope r is necessarily parallel to K in $N(K)$, and so in particular if $\partial V = S^3$ then r is an integral slope. Conversely a manifold resulting from surgery on a framed link $\mathbf{L} = (L; r_1, \ldots, r_n)$ in S^3 is the boundary of the 4-manifold $V(\mathbf{L})$ obtained by attaching 2-handles to the 4-ball along $N(L)$ using attaching maps determined by the (integral) framing slopes. The attaching map associated to the framing $p \in \mathbf{Z}$ of a knot K is required to send the curves $S^1 \times \{*\}$ to curves of slope p on $\partial N(K)$. Thus as a corollary of Lickorish's approach to Theorem 3.1, we obtain a proof that the 3-dimensional oriented cobordism group is zero.

The main result of Kirby [63], interpreted in terms of the rational calculus, may be stated as follows.

THEOREM 3.2 (Kirby [63]). *Let* $\mathbf{L} = (L; r_1, \dots, r_n)$ *and* $\mathbf{L}' = (L'; r'_1, \dots, r'_{n'})$ *be two sets of surgery data such that* $L(r_1, r_2, \dots, r_n) \cong L'(r'_1, \dots, r'_{n'})$ *by an orientation preserving homeomorphism. Then one can obtain* $(L'; r'_1, \dots, r'_{n'})$ *from* $(L; r_1, \dots, r_n)$ *by a finite sequence of rational calculus moves.*

PROOF. We shall denote by "$\sim$" the equivalence relation on the set of surgery presentations generated by the three moves of the rational calculus. The heart of the proof is an application of the Cerf theory of functions on a smooth manifold, though a certain amount of work must first be done to reach this point.

To begin with, it is elementary to show that any surgery presentation can be modified by the moves listed above to produce a framed link [86]. Hence we may suppose that $\mathbf{L}$ and $\mathbf{L}'$ are in fact framed links. The next step is to show that there are framed links $\mathbf{L}_1, \mathbf{L}'_1$ for which $\mathbf{L} \sim \mathbf{L}_1$, $\mathbf{L}' \sim \mathbf{L}'_1$, and $V(\mathbf{L}_1) = V(\mathbf{L}'_1)$. Kirby accomplished this as follows. Recall that $\partial V(\mathbf{L}) = L(r_1, r_2, \dots, r_n)$ and $\partial V(\mathbf{L}') = L'(r'_1, \dots, r'_{n'})$, so that by hypothesis we may form $U = V(\mathbf{L}) \cup (-V(\mathbf{L}'))$ by identifying boundaries. The signature of U is then arranged to be zero by adding several ± 1-framed trivial knots to $\mathbf{L}$ (this will not change its "$\sim$" class) and hence U bounds a compact, connected, orientable, smooth 5-manifold X [94]. Thinking of X as a cobordism between $V(\mathbf{L})$ and $V(\mathbf{L}')$, we can apply standard handle-swapping arguments to produce an X built on $V(\mathbf{L})$ with only 2-handles and 3-handles. The level V of X, lying between the 2-handles and the 3-handles, can now be seen to be a connected sum of $V(\mathbf{L})$ with several copies of $S^2 \times S^2$ and $S^2 \tilde{\times} S^2$ (the non-trivial S^2-bundle over S^2). Kirby shows that these connected sums can be realized by his versions of the "$\sim$" moves on $\mathbf{L}$, and so without loss of generality we may suppose that $V(\mathbf{L}) = V$. Consideration of the dual handle decomposition of X shows that we may also suppose that $V(\mathbf{L}') = V$.

To complete the proof, we now observe that the two handle decompositions of V coming from the identifications $V(\mathbf{L}) = V = V(\mathbf{L}')$ give rise to two Morse functions f, f' each of which has exactly one critical point of index 0 and none of index 1, 3, 4. Following the method of J. Cerf, there is a homotopy $t \mapsto f_t$ where $f_0 = f$, $f_1 = f'$, and f_t is a Morse function with distinct critical values for all but finitely many values $t_1, t_2, \dots, t_n$ of t. At the exceptional values, the functions f_{t_i} are of one of two types denoted by α, β. The α-type functions are Morse with distinct critical values except for one critical point where a certain "birth or death" phenomenon occurs. The β-type are Morse with distinct critical values except for two which coincide. Kirby shows that after a further sequence of "$\sim$" moves on $\mathbf{L}$ and $\mathbf{L}'$, a homotopy between the new f and f' may be found for which each f_t is Morse (so no α-type functions arise), has a unique critical point of index 0, and none of index 1, 3, 4. Away from a finite set $\{s_1, s_2, \dots, s_m\} \subset (0, 1)$, the descending 2-disks of the index 2 critical points of f_t are disjoint and intersect $S^3 = \partial B^4$ in a link L_t which underlies $\mathbf{L}_t$, a framed link presentation of ∂V. As t varies in $[0, 1] \setminus \{s_1, s_2, \dots, s_m\}$, the effect on $\mathbf{L}_t$ is an isotopy. When $t = s_j$, the descending 2-disk of some index 2 critical point of f_{s_j} strikes an index 2 critical point of f_{s_j} of smaller critical value. This corresponds to sliding one descending 2-disk over the other and so as t passes through s_j, $\mathbf{L}_t$ changes by what Kirby calls a handle slide (one of his two basic moves). Finally Fenn and Rourke [30] showed that handle slides may be accomplished by a sequence "$\sim$" moves, thus completing the proof. □

Theorem 3.2 should not be viewed as providing a classification, or even an effective enumeration, of 3-dimensional manifolds. Rather, it relates framed links in much the same way that finite group presentations are related by the Tietze transformations. It has been effective as a tool for producing examples and unexpected homeomorphisms. Furthermore, one can often effectively express, or even define, invariants of a given 3-manifold in terms of framed link data. An important example of the former, due to C. Lescop, may be found in [67], and of the latter, due to R. Kirby and P. Melvin, in [65].

4. Surfaces and filling

Surfaces have played an important role in our understanding of the topology of 3-dimensional manifolds. Thus it is not surprising that we devote a fair amount of attention to their relationship with fillings. We shall assume below that M is a compact, connected, orientable, irrededucible, ∂-irreducible 3-manifold whose boundary contains a torus T. All surfaces in M will be assumed to be smooth and either properly embedded or contained in ∂M.

4.1. *Destroying an incompressible surface by Dehn filling*

Manifolds which contain essential surfaces are subject to strong structure theorems and so it is natural to examine how a closed, incompressible surface $S \subset M$ can compress in some Dehn filling $M(T; r)$ of M. It is easy to construct examples where S does compress. For instance the exterior of a satellite knot $K \subset S^3$ always contains an essential torus which clearly compresses in $M(\mu_K) = S^3$. Thus a better strategy is to determine what sort of relationship exists between two slopes r_1, r_2 for which S compresses in $M(T; r_1)$ and $M(T; r_2)$. To examine this problem, let us consider M_1, the closure in M of the component $M \setminus S$ which contains T. If S separates M, then there is a unique surface $S_1 \subset \partial M_1$ corresponding to S, otherwise there are two such surfaces S_1, S_2. Since S compresses in a filling $M(T; r)$ if and only if some S_i compresses in $M_1(T; r)$, our original problem translates into the study of the ∂-reducibility of the filled manifolds $M_1(T; r)$.

The most obvious examples of ∂-reducibility fillings occur when $M_1 \cong T \times I$, for then $M_1(T; r) \cong S^1 \times D^2$ for each slope r. But even if we assume that $M_1 \ncong T \times I$, there could still be a properly embedded annulus $A \subset M_1$ whose boundary consists of an essential curve on S_1, and an essential curve of some slope r_0 on T. Clearly S_1 compresses in $M_1(T; r_0)$ and if r is another slope on T for which $M_1(T; r)$ is ∂-reducible, then $M_1(T; r')$ is ∂-reducible for all slopes r' on T for which $\Delta(r, r_0) = \Delta(r', r_0)$. This is because one may perform a series of Dehn twists along the annulus A to construct a homeomorphism of M_1 which leaves the boundary components of M_1 invariant and which takes the slope r to r'.

EXAMPLE 4.1. Let V be a standard unknotted solid torus in S^3 and let $V_0 \subset \text{int}(V)$ be a tubular neighbourhood of the core of V. Suppose that $|q| \geqslant 2$ and $K \subset \partial V_0$ is a (p,q) torus knot with exterior V_K. The manifold V_K is irreducible and ∂-irreducible. Since ∂V_0 is parallel to ∂V in V, there is a properly embedded annulus in V_K whose boundary consists of a (p,q) torus knot on ∂V and a curve of slope r_0, say, on $\partial N(K)$. As $V_K(\mu_K) \cong S^1 \times D^2$ and $\Delta(\mu_K, r_0) = 1$, the remarks above imply that $V_K(r) \cong S^1 \times D^2$ for all slopes r satisfying $\Delta(r, r_0) = 1$. We shall see in Section 5.2 that $V_K(r_0) \cong (S^1 \times D^2)\#L(p,q)$ and $V_K(r)$ is irreducible and ∂-irreducible whenever $\Delta(r, r_0) \geqslant 2$.

DEFINITION 4.2. A (p,q) *cable space* is any manifold homeomorphic to the exterior of a (p,q) torus knot standardly embedded in the interior of an unknotted solid torus.

Cable spaces admit homeomorphisms which exchange their two boundary components and so from Example 4.1 we see that if T is a boundary component of a cable space C of type (p,q), then there is a slope r_0 on T, called the *cabling slope*, for which

$$C(T;r) \cong \begin{cases} S^1 \times D^2\#L(q,p) & \text{if } r = r_0, \\ S^1 \times D^2 & \text{if } \Delta(r, r_0) = 1, \end{cases}$$

and is irreducible and ∂-irreducible otherwise. Hence $C(T;r)$ is ∂-reducible if and only if $\Delta(r, r_0) \leqslant 1$.

Thus the presence of an essential annulus can drastically affect the ∂-reducibility properties of the manifolds $M_1(T;r)$. Happily, if there is no such annulus then the situation is quite controllable. The following theorem was originally conjectured by C.McA. Gordon and J. Luecke, who verified it in many instances [21]. Y.-Q. Wu provided a completed proof.

THEOREM 4.3 (Wu [108]). *Let M be a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold whose boundary contains a torus T. Suppose further that $M \not\cong T \times I$. If r_1 and r_2 are two slopes on T for which $M(T;r_1)$ and $M(T;r_2)$ are ∂-reducible, then either*

(i) $\Delta(r_1, r_2) \leqslant 1$, *or*

(ii) $\Delta(r_1, r_2) > 1$ *and there are a component $S \neq T$ of ∂M and an annulus A properly embedded in M such that*

- *∂A consists of an essential curve on S and an essential curve $C_0 \subset T$;*
- *if r_0 denotes the slope of C_0 on T, then $M(T;r)$ is ∂-reducible if and only if $\Delta(r_0, r) \leqslant 1$;*
- *$\partial M \setminus (T \cup S)$ is incompressible in $M(T;r)$ for all $r \neq r_0$.*

REMARK 4.4. Example 4.1 is complemented by an example of a knot in $S^1 \times D^2$, due to Gabai [36] and Berge [6], whose exterior contains no essential annulus, but which has three slopes whose associated fillings yield $S^1 \times D^2$, the maximal number of ∂-reducible fillings permitted by Theorem 4.3 (cf. Table 1).

The proof of Wu's theorem is based on the combinatorial/topological theory of intersection graphs which originated in the work of Litherland [70] and was later developed into a powerful tool in a series of papers by Gordon and Luecke (see, e.g., [21,48,49]). As this theory is of paramount importance in the study of the Dehn filling operation, we shall take some pains to describe it in the context of Theorem 4.3. An excellent introduction to this subject may be found in [44].

Choose properly embedded compressing disks $D_1 \subset M(T;r_1)$ and $D_2 \subset M(T;r_2)$ whose boundaries are essential curves lying in components S_1 and S_2 (possibly equal) of $\partial M \setminus T$. After an isotopy we may assume that for $j = 1,2$, the disk D_j intersects the filling solid torus V_j in a collection of n_j meridian disks. The incompressibility of S_j in M implies that $n_j > 0$. Assume that D_j has been chosen, amongst all such compressing disks, so as to minimize n_j.

Consider the planar surfaces $P_j = D_j \cap M$, isotoped so that they are transverse to each other and so that they intersect minimally on T. Then each component of $\partial_0 D_1 = T \cap \partial P_1$ meets each component of $\partial_0 D_2 = T \cap \partial P_2$ in $\Delta(r_1, r_2)$ points. Number the components of $\partial_0 D_1$ so that as we travel along any component of $\partial_0 D_2$ (in some direction) the indices of the components of $\partial_0 D_1$ we encounter are $1, 2, \ldots, n_1, 1, 2, \ldots, n_1, \ldots$ (repeated $\Delta(r_1, r_2)$ times). Order the components of $\partial_0 D_2$ similarly.

Observe that $P_1 \cap P_2$ is a finite collection of arcs and circles and by our assumptions on P_1 and P_2 it follows that no component of $P_1 \cap P_2$ is an inessential curve in either P_1 or P_2. For $j = 1,2$ we define the *labelled graph* $\Gamma_j \subset D_j$ of this configuration to be the graph in D_j whose (fat) vertices are the disks $D_j \cap V_1$ and whose edges are the arc components of $P_1 \cap P_2$. We label an endpoint of an edge of Γ_1 (Γ_2) which lies in the interior of D_1 (D_2) with the number of the corresponding component of $\partial_0 D_2$ ($\partial_0 D_1$).

Two vertices of Γ_j are said to be *parallel* if when their boundaries are given the orientations induced from some orientation of P_j they are homologous on T. Otherwise they are called *antiparallel*. Choose signs $+, -$ for each vertex of Γ_j so that parallel vertices have the same sign and antiparallel ones have opposite ones (see Figure 4). The orientability of M, P_1, and P_2 implies the following useful rule.

LEMMA 4.5 (The parity rule). *An arc component of $P_1 \cap P_2$ connects vertices of the same sign in Γ_1 if and only if it connects vertices of opposite signs in Γ_2.*

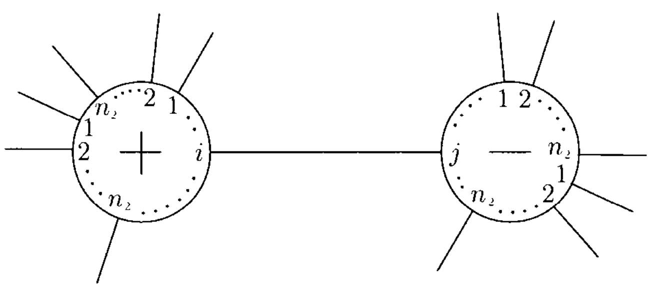


Fig. 4. An edge of the intersection graph Γ_1.

The philosophy of this method is that under certain hypotheses, for instance that $\Delta(r_1, r_2)$ be relatively large, combinatorial arguments can be used to show that Γ_1 or Γ_2 contain configurations of a specified type. Then topological arguments imply that if Γ_1 or Γ_2 contain such configurations, useful topological information about $M(r_2)$ or $M(r_1)$ can be deduced. As an example of such a configuration we introduce the following notion.

DEFINITION 4.6. A *Scharlemann cycle* in Γ ($\Gamma = \Gamma_1$ or Γ_2) is a subgraph C of Γ which satisfies

- C bounds a disk face $\mathcal{D}$ of Γ;
- all the vertices of C are parallel;
- there is a (mod n) label pair $(i, i+1)$ which gives the labels of the endpoints of each edge of C.

Suppose for the moment that Γ_2 contains a Scharlemann cycle with disk face $\mathcal{D} \subset D_2$ and distinguished label pair $(i, i+1)$. Let F be a subdisk of D_1 which contains the edges of C. Since i and $i+1$ are successive labels of Γ_1, F slices off a product $H \cong D^2 \times I$ from the filling torus of $M(T; r_1)$ (see Figure 5).

If C has $p > 1$ edges, then the 3-manifold defined by

$$N(F, \mathcal{D}) = N(F \cup H \cup \mathcal{D}) \subset M(T; r_1)$$

may be described as the union of a 0-handle $N(F)$, a 1-handle H, and a 2-handle $N(\mathcal{D})$ whose attaching circle which runs over the 1-handle p times, always in the same direction (because the vertices of C are all of the same sign). Thus $N(F, \mathcal{D})$ is a punctured lens space of order p contained in $M(T; r_1)$. In particular $\partial N(F, \mathcal{D}) \cong S^2$. After a small isotopy we may arrange for $\partial N(F, \mathcal{D}) \cap D_1 = F$. Set

$$D_1' = (D_1 \setminus F) \cup (\overline{\partial N(F, \mathcal{D}) \setminus F}).$$

By construction D_1' is a disk with the same boundary as D_1, but which intersects the filling solid torus V_1 in $n_1 - 2$ meridian disks (the two disks of intersection of V_1 corresponding to

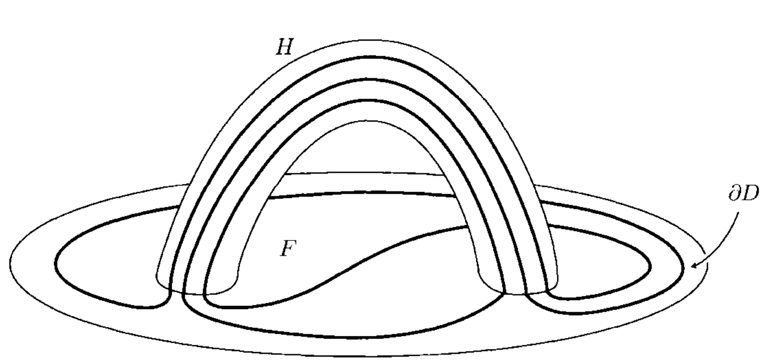


Fig. 5. The topology of a Scharlemann cycle.

vertices of indices $i, i+1$ have been eliminated). But this is impossible by the minimality of n_1. Thus we conclude that Γ_2 cannot contain a Scharlemann cycle. A similar conclusion holds for Γ_1.

PROOF OF THEOREM 4.3. The idea is to parlay the non-existence of Scharlemann cycles in Γ_1 and Γ_2 into constraints on distance.

We observe first of all that there is no vertex x of Γ_1 such that for each vertex y of Γ_2 there is an edge $e(y)$ of Γ_2 incident to y with label x, and which connects y to a parallel vertex of Γ_2. Otherwise it would be possible to construct an oriented subgraph Λ of Γ_2, homeomorphic to a circle, such that the leading vertex of each edge of Λ is labelled x and each vertex of Γ_2 which lies in the disk bounded by Λ is parallel to the vertices of Λ. It turns out that such a graph, called a *great x-cycle*, always contains a Scharlemann cycle of Γ_2 (Lemma 2.6.2 of [21]), contradicting our choice of D_1.

Applying the parity rule (Lemma 4.5) we may therefore assume that for each vertex x of Γ_1, there is a vertex $y(x)$ of Γ_2 such that each edge of Γ_1 incident to x with label $y(x)$ connects x to a parallel vertex of Γ_1.

As a first case, suppose that $S_1 \neq S_2$, so that the graphs Γ_1, Γ_2 lie in the interiors of D_1, D_2. For each vertex x of Γ_1 choose a vertex $y(x)$ of Γ_2 as described in the previous paragraph and consider the subgraph $\Lambda \subseteq \Gamma_1$ defined as the union over all vertices x of Γ_1 of the edges of Γ_1 incident to x with label $y(x)$. Let Λ_0 be an innermost component of Λ and observe that if we assume that $\Delta(r_1, r_2) \geqslant 2$, then each vertex of Λ_0 has valency at least 2. The outer frontier of Λ_0 therefore contains a subgraph homeomorphic to a circle. As Λ_0 is innermost, all the vertices of Γ_1 contained in the disk $E \subset D_1$ bounded by this circle are parallel. The argument now proceeds by showing that for some vertex y of Γ_2 there is a great y-cycle of Γ_2 contained in Λ_0. Thus as in the previous case there is a Scharlemann cycle based on the label y contained in Γ_2, contrary to our choice of D_1. We must therefore have $\Delta(r_1, r_2) \leqslant 1$ when $S_1 \neq S_2$ (actually all we used was that $\partial_0 D_1 \cap \partial_0 D_2 = \emptyset$).

The proof of the final case where $S_1 = S_2$ proceeds along similar lines, but is complicated by the fact that now there may be *boundary edges* of Γ_1 and Γ_2, i.e., edges which have at least one end on $\partial D_1 \cap \partial D_2$. The argument of the previous paragraph implies that if $\Delta(r_1, r_2) > 1$ then every vertex of Γ_2 is incident to ∂D_2 (Lemma 2.6.4 of [21]). When $\Delta(r_1, r_2) \geqslant 3$, Gordon and Luecke showed that there are a pair of boundary edges which are parallel in both Γ_1 and Γ_2 (Proposition 2.5.8 of [21]). An annulus can then be built from these parallelisms and it is fairly easy to see that it satisfies the properties stated in the theorem. The final ingredients of the proof are due to Wu [108]. He showed that (M, T) could be put in a certain standard position with respect to D_1 and then uses W. Jaco's 2-handle addition theorem [61] to argue that if there is no annulus of the desired sort in (M, T), then $\Delta(r_1, r_2) = 1$. □

Intersection graphs can be defined for any pair of surfaces $\widehat{P}_1 \subset M(T; r_1)$ and $\widehat{P}_2 \subset M(T; r_2)$, though if they are to contain interesting information we must assume that the arcs of Γ_1 and Γ_2 are essential in both P_1 and P_2. This cannot always be accomplished without additional constraints on the surfaces. Consider the hypotheses

(H1) $\widehat{P}_j$ is essential and M contains no essential surface homeomorphic to it;

(H2) $\widehat{P}_j$ is a Heegaard surface and the core of the filling solid torus is not isotopic into it.

Standard arguments show that if both $\widehat{P}_1$ and $\widehat{P}_2$ satisfy (H1) and are chosen to minimize n_1 and n_2, then the arcs in the intersection graphs are essential. If one of the surfaces satisfies (H1) and the other (H2) this fact follows as in Gabai's paper [34], while if both satisfy (H2) the result was proved by Rieck [83].

Scharlemann cycles are defined as above for general intersection graphs and their existence provides good topological information. We record the following proposition for later use.

PROPOSITION 4.7. *Suppose that the arcs of Γ_1 and Γ_2 are essential in both P_1 and P_2. Let D be a Scharlemann cycle in Γ_1 which has exactly p edges, all of which lie in a disk $F \subset \widehat{P}_2$. Then $M(T; r_2)$ contains a once-punctured lens space of order p. Furthermore if $\widehat{P}_2$ is chosen so as to minimize n_2 amongst all surfaces in $M(T; r_2)$ of a given homeomorphism type which satisfy* (H1), *it is a 2-sphere which bounds a punctured lens space in $M(T; r_2)$.*

PROOF. We need only verify the second conclusion. But if $\widehat{P}_2 \not\cong S^2$ satisfies (H1) and minimizes n_2, then $\widehat{P}'_2 = \overline{\partial N(\widehat{P}_2, \mathcal{D}) \setminus \widehat{P}_2} \cong \widehat{P}_2$ is also essential and has $n'_2 < n_2$, contrary to our construction. On the other hand, if $\widehat{P}_2 \cong S^2$, then the minimality of n_2 implies that $\widehat{P}'_2$ is a 2-sphere which bounds a 3-ball $B \subset M(T; r_2)$. As B cannot contain $\widehat{P}_2$, it follows from the construction that $\widehat{P}_2$ bounds a punctured lens space in $M(T; r_2)$. □

4.2. *Incompressible surfaces which survive Dehn filling*

Let us turn the question of the previous section around and ask what conditions on an incompressible surface guarantee that it survives most fillings. A simple criterion, due to W. Menasco, has proven useful in understanding the manifolds resulting from Dehn surgery on alternating knots [73], star knots [80], 2-bridge knots [26], and arborescent knots [109].

THEOREM 4.8 (Menasco [73]). *Let T be a toral boundary component of an irreducible, orientable 3-manifold M, and suppose that $S \subseteq \partial M \setminus T$ is incompressible. If there are two disjoint, essential annuli A_1, A_2 properly embedded in M, each running between S and T, and if the curves $A_1 \cap S$, $A_2 \cap S$ are not isotopic on S, then S is incompressible in $M(T; r)$ for each slope r on T other than the common slope of the curves $A_1 \cap T$, $A_2 \cap T$.*

PROOF. Let r_0 be the slope on T of the two annuli and suppose that D is a compressing disk for S in $M(T; r)$, $r \neq r_0$. Standard arguments imply that D can be chosen transverse to $A_1 \cup A_2$ so that each arc of $D \cap A_j$ runs from one boundary component of A_j to the other. The edges around each vertex of the intersection graph Γ_D formed from $D \cap (A_1 \cup A_2)$ alternate between A_1 edges and A_2 edges. An outermost argument implies that there is a rectangular face R of Γ_D whose boundary consists of four arcs: an edge of A_1, an edge of A_2, an arc β in ∂D, and an arc in $\partial(D \cap M) \setminus \partial D$. If α_j is the circle $A_j \cap S$, then there

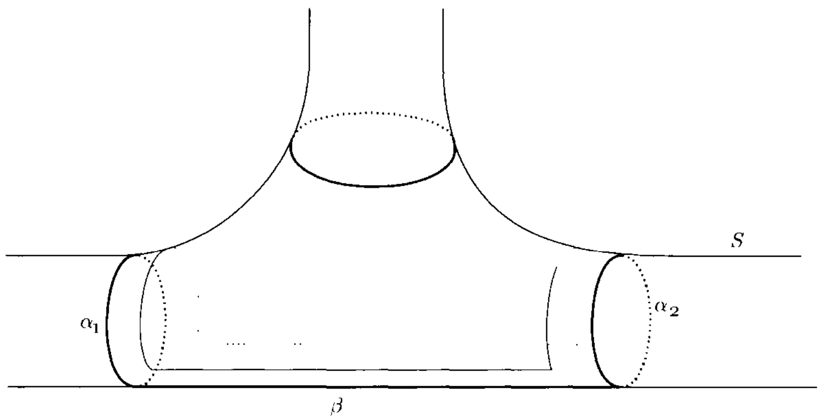


Fig. 6.

is a component of the boundary of a regular neighbourhood of $\alpha_1 \cup \alpha_2 \cup \beta$ which is an essential curve on S (because α_1 and α_2 are not isotopic on S) and which compresses in M (by the existence of R), contradicting the incompressibility of S (see Figure 6). Thus S cannot compress in $M(T; r)$ when $r \neq r_0$. □

Another useful criterion which guarantees incompressibility after filling is due to D. Gabai. Before stating it we must introduce several new notions.

DEFINITION 4.9. Let N be a codimension 0 submanifold of ∂M. An oriented, properly embedded surface $(S = \bigcup_{i=1}^m S_i, \partial S) \subset (M, N)$, S_i connected, is called *(Thurston) norm minimizing* in $H_2(M, N)$ if it is incompressible and

$$\max\left\{0, \sum_{i=1}^{m} -\chi(S_i)\right\} \leqslant \max\left\{0, \sum_{j=1}^{n} -\chi(S'_j)\right\}$$

for each oriented, properly embedded surface $S' = \bigcup_{j=1}^n S'_j \subset M$ which carries the same homology class as S.

Thus norm minimizing surfaces provide representatives of minimal complexity for a given homology class. Interest in them was sparked by work of Thurston [98] in which he showed that compact leaves of nice 2-dimensional foliations are norm minimizing. Gabai studied the converse problem: when is a norm minimizing surface a union of leaves of a nice foliation on M? An interesting example of his work is the following theorem.

THEOREM 4.10 (Gabai [33]). *Let M be a compact, connected, orientable, irreducible 3-manifold which contains no essential tori. Assume further that ∂M is a torus and $H_2(M) \neq 0$. Suppose that $S \subset \mathrm{int}(M)$ is a non-separating, closed, Thurston norm minimizing surface. Then for all slopes r on ∂M, with at most one exception, S is incompressible in $M(r)$, indeed norm minimizing, and $M(r)$ is irreducible.*

PROOF. The result is proved by showing that there is a slope r_0 on T such that for any other slope r, there is a nicely behaved 2-dimensional foliation on $M(r)$ for which S is a leaf. Hence Thurston's result [98] mentioned above implies that S is norm minimizing in $M(r)$. Using work of Rosenberg [87] it follows that the special properties of the foliation, which we will not discuss here, also imply that $M(r)$ is irreducible.

The foliation is constructed by induction along an appropriately chosen *sutured manifold hierarchy* of $M(r)$. A *sutured manifold* is a pair (N,γ) where N is a compact, oriented 3-manifold and $\gamma \subset \partial N$ is a closed 1-submanifold which splits ∂N into the union of two subsurfaces $\overline{R}_+(\gamma), \overline{R}_-(\gamma)$ oriented so that $\partial N = \overline{R}_+(\gamma) \cup_\gamma (-\overline{R}_-(\gamma))$. Let $A(\gamma) \subset \partial N$ be a collection of disjoint annuli which forms a regular neighbourhood of γ and let $T(\gamma)$ be a collection of toral components of ∂N. Set $R_\pm(\gamma) = \overline{R}_\pm(\gamma) \setminus (\mathrm{int}(A(\gamma)) \cup T(\gamma))$. The goal is to construct 2-dimensional foliations on N which are tangent to $R_\pm(\gamma)$ and transverse to $A(\gamma) \cup T(\gamma)$. This is not always possible of course because if it does occur, Thurston's work mentioned above implies that both R_+ and R_- would be norm minimizing.

In our situation, M is successively cut open along oriented surfaces $S = S_1, S_2, \dots, S_n$ disjoint from T to construct a sequence of sutured manifolds

$$(M,\emptyset) \xrightarrow{S} (M_1,\emptyset) \xrightarrow{S_2} (M_2,\gamma_2) \xrightarrow{S_3} \cdots \xrightarrow{S_n} (M_n,\gamma_n),$$

where the sutures $\gamma_2,\dots,\gamma_n$ are essential curves formed from pieces of ∂S_i by paying strict attention to the orientations of the decomposing surfaces S_i. Using an involved notion of sutured manifold complexity, Gabai shows that $S_2,\dots,S_n$ can be chosen in such a way that the surfaces $R_\pm(\gamma_i)$ are norm minimizing and (M_n,γ_n) splits as a disjoint union

$$(M_n,\gamma_n) = \big(F \times I, \partial F \times \{1/2\}\big) \cup (H,\gamma'),$$

where $T \subset H \cong T \times I$ and $\emptyset \neq \gamma' \subset \partial H \setminus T$. Let r_0 be the slope on T corresponding to γ', through the product structure on H.

For each value of i, $T \subset M_i$ and so for any slope r on T we can form a new sequence

$$\big(M(r),\emptyset\big) \xrightarrow{S} \big(M_1(T;r),\emptyset\big) \xrightarrow{S_2} \big(M_2(T;r),\gamma_2\big) \xrightarrow{S_3} \cdots \xrightarrow{S_n} \big(M_n(T;r),\gamma_n\big).$$

Now from above $(M_n(T;r),\gamma_n) = (F \times I, \partial F \times \{1/2\}) \cup (H(T;r),\gamma')$ where $H(T;r)$ is a solid torus whose meridian slope, r' say, corresponds to r through the product structure on H. If $r \neq r_0$, then the components of γ' do not represent r', and cutting $H(T;r)$ open along an appropriately chosen meridian disk D then gives a final decomposition

$$\begin{aligned}\big(M_n(T;r),\gamma_n\big) &\xrightarrow{D} (M_{n+1},\gamma_{n+1})\\ &= \big(F \times I, \partial F \times \{1/2\}\big) \cup \big(D^2 \times I, \partial D^2 \times \{1/2\}\big).\end{aligned}$$

One now proceeds, using a delicate argument, to show that starting with the product structure on (M_{n+1},γ_{n+1}), nice foliations may be inductively constructed on $(M_n(T;r),\gamma_n)$, $\dots$, $(M_2(T;r),\gamma_2)$, $(M_1(T;r),\emptyset)$, $(M(r),\emptyset)$ in such a way that S becomes a leaf of the last one, thus completing the proof. □

4.3. *Fillings parameterized by boundary slopes*

If $(F, \partial F) \subset (M, T)$ is an essential surface with non-empty boundary of slope r, then F may be capped off to form a closed surface $\widehat{F} \subset M(T; r)$ by attaching meridional disks from the filling solid torus. The question that we address here is whether or not $\widehat{F}$ is essential in $M(T; r)$. It is easy to see that the answer is no in general – just perform meridional filling on the exterior of a knot whose meridian is a boundary slope (e.g., a product knot). Nevertheless, we shall see that if the surface is chosen with some care, then the answer is often yes. An example of this which has important applications to surgery on knots in the 3-sphere was provided by Gabai [34]. Recall that a Seifert surface of a knot $K \subset S^3$ is a properly embedded, compact, connected, orientable surface $F \subset M_K$ whose boundary is a longitude of K.

THEOREM 4.11 (Gabai [34]). *If F is a minimal genus Seifert surface for a knot in S^3, then $\widehat{F}$ is essential in $M_K(0)$.*

PROOF. A Seifert surface is of minimal genus if and only if it minimizes the Thurston norm of its class in $H_2(M_K, \partial M_K)$. Gabai uses sutured manifold theory (see the proof of Theorem 4.10) to show that M_K supports a nice 2-dimensional foliation $\mathcal{F}$ for which F is a leaf and $\mathcal{F}|\partial M_K$ is a foliation by longitudinal circles. The result then follows from Thurston [98] by observing that $\mathcal{F}$ naturally extends to a foliation of $M_K(0)$ in which $\widehat{F}$ is a leaf. □

As a corollary of this result, Gabai deduced that non-trivial knots in the 3-sphere satisfy the property R conjecture, i.e., that no surgery on a non-trivial knot in S^3 can yield $S^1 \times S^2$.

COROLLARY 4.12. *If $K \subset S^3$ is a non-trivial knot, then $M_K(0) \ncong S^1 \times S^2$.*

Suppose now that r is a boundary slope on T. The proof of the next theorem shows that though there are potentially many different essential surfaces F having non-empty boundary of slope r, choosing one with a minimal number of boundary components often results in $\widehat{F}$ being essential.

Let $b_1(M)$ denote the first Betti number of M.

THEOREM 4.13 (Culler, Gordon, Luecke and Shalen [21]). *Let M be a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold with boundary a torus T and $b_1(M) = 1$. If r is a boundary slope on T, then one of the following four possibilities occurs.*

(i) *$M(r)$ is Haken.*
(ii) *$M(r)$ is the connected sum of two non-trivial lens spaces.*
(iii) *There is a closed essential surface $S \subset M$ which compresses in $M(r)$ but which remains incompressible in $M(r')$ as long as $\Delta(r, r') > 1$.*
(iv) *M is a planar surface bundle over the circle (so $M(r) \cong S^1 \times S^2$).*

Furthermore, in the last case, if r is the boundary slope of an essential surface which is not a fibre in any fibration of M over the circle, then (iii) *also holds.*

PROOF. Let F be a properly embedded, separating, essential surface in M of boundary slope r such that each component of F has non-empty boundary and such that n, the number of components of ∂F, is minimal subject to these conditions. Then F is either connected or has exactly two components, each of which is non-separating. If the latter arises we may assume that F consists of two parallel copies of a surface G. The assumption that $b_1(M) = 1$ can be used to deduce that if G is given any orientation, then all the boundary components of G are parallel on T.

Denote by $\widehat{F}$ the closed surface in $M(r)$ obtained by attaching meridian disks to F. The proof will show that either $\widehat{F}$ is an incompressible surface of positive genus and conclusion (i) of the theorem holds, or $\widehat{F}$ is a separating 2-sphere and (ii) holds, or $\widehat{F}$ is a non-separating 2-sphere and (iv) holds, or M contains a closed, essential surface disjoint from $\widehat{F}$ which satisfies the conclusions of (iii). For simplicity we shall stick to the case where F is connected and of positive genus. The other cases are handled similarly.

The surface F splits M into two manifolds X and X'. Using the minimality of n, a collection of $(n-1)$-disjoint 2-handles in $(X, \partial X)$ may be found which, when attached to a collar of ∂X in X, produces a compression body $W \subset X$ with outer boundary ∂X and inner boundary a connected surface S of genus equal to genus$(F) + 1$. It turns out that S always compresses in $M(r)$.

Suppose first of all that S is incompressible in X. We will prove that conclusion (iii) of the theorem holds.

Now S must be essential in M, for otherwise it would cobound a copy of $T \times I$ with T, and hence F would be homotopic into T, contrary to hypothesis. Suppose that S compresses in $M(r')$ for some slope r' on T such that $\Delta(r, r') > 1$. Since $b_1(M) = 1$, S separates M, and hence it cobounds with T an irreducible, ∂-irreducible submanifold $N \not\cong T \times I$ of M. Clearly S compresses in both $N(T; r)$ and $N(T; r')$ and so by Theorem 4.3 there is an annulus A properly embedded in N whose boundary consists of an essential curve on S and an essential curve C_0 of slope r_0, say, on T. Furthermore $\Delta(r, r_0) = \Delta(r', r_0) = 1$, so that in particular $r \neq r_0$. Now make A transverse to F in such a way that ∂F intersects C_0 minimally. If $F \cap A \neq \emptyset$, choose an arc of intersection which is outermost in A. The ∂-irreducibility of F implies that this arc can be removed from $F \cap A$ by an isotopy, contrary to minimality of $F \cap C_0$. Thus $\partial F \cap C_0 = \emptyset$, which is impossible as $r \neq r_0$. Hence S must remain incompressible in $M(r')$ for every slope r' for which $\Delta(r, r') > 1$ (so conclusion (iii) of the theorem holds).

Similar constructions can be performed on X' to produce a closed surface $S' \subset X'$ (homeomorphic to S) which, if incompressible in M, can be shown to compress in $M(r)$ and to be incompressible in $M(r')$ whenever $\Delta(r, r') > 1$. Hence if conclusion (iii) of the theorem does not hold, both S and S' compress in M. The filling solid torus in $M(r)$ is cut into pieces by $\widehat{F}$, each a 3-ball, which are attached successively to X and X' along annuli in T of slope r. Hence we have $M(r) = \widehat{X} \cup \widehat{X}'$ where $\widehat{X}$ $(\widehat{X}')$ is obtained from X (X') by attaching 2-handles. One can now combine Jaco's 2-handle addition theorem [61] with the particular properties of the construction to show that $\widehat{X}$ is irreducible and its boundary $\widehat{F}$ incompressible. A similar conclusion holds for $\widehat{X}'$. Thus $M(r) = \widehat{X} \cup_{\widehat{F}} \widehat{X}'$ is irreducible and contains $\widehat{F}$, a non-planar, incompressible surface. Conclusion (i) of the theorem therefore holds. □

4.4. *Creating essential surfaces of non-negative Euler characteristic*

In this section we continue our discussion of the surfaces $\widehat{F} \subset M(T; r)$ formed by attaching meridian disks to surfaces in M which intersect T in a collection of curves of slope r.

Suppose that M contains no essential surface homeomorphic to some fixed surface $\widehat{F}$ of non-negative Euler characteristic, while $M(T; r_1)$ and $M(T; r_2)$ do. How are the slopes r_1, r_2 related? Taking $\widehat{F} \cong D^2$ we see that this line of enquiry generalizes the work of Section 4.1. Questions of this nature are of interest because as we shall see in Section 5, the structure of the fillings $M(T; r)$ is determined to a great extent by the (non) existence of such surfaces. The next three theorems are typical of the results available.

THEOREM 4.14 (Gordon and Luecke [50]). *Suppose that M is an orientable, irreducible, ∂-irreducible 3-manifold and T is a toral boundary component of M. If $M(r)$ and $M(r')$ are reducible 3-manifolds, then $\Delta(r, r') \leqslant 1$.*

THEOREM 4.15 (Oh [81], Wu [110]). *Suppose that M is an orientable, irreducible, ∂-irreducible 3-manifold and T is a toral boundary component of M. Suppose further that M contains no essential tori or annuli. If $M(r)$ is reducible and $M(r')$ contains an essential torus, then $\Delta(r, r') \leqslant 3$.*

Before stating the last of the three theorems, we introduce the Whitehead link.

EXAMPLE 4.16 (*The Whitehead link*). The *Whitehead link* is the two-component link $L \subset S^3$ pictured in Figure 7. Set $W = M_L$ and fix a component T_0 of ∂W. There is an isotopy of S^3 exchanging the two components of L and so it is immaterial which of them we choose as T_0. As $L \subset S^3$ we may identify the slopes on T_0 canonically with $P^1(\mathbf{Q})$.

THEOREM 4.17 (Gordon [45]). *Let M be a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold which contains no essential tori and whose boundary contains a torus T. Further let W be the Whitehead link exterior and T_0 a boundary component of W. If $M(T; r_1)$ and $M(T; r_2)$ contain essential tori, then either*

(i) $\Delta(r_1, r_2) \leqslant 5$; *or*

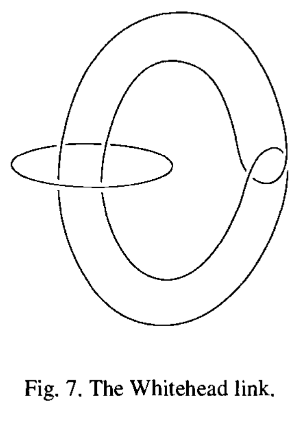


Fig. 7. The Whitehead link.

(ii) $\Delta(r_1, r_2) = 6$ *and* $M \cong W(T_0; -2)$; *or*
(iii) $\Delta(r_1, r_2) = 7$ *and* $M \cong W(T_0; 5/2)$; *or*
(iv) $\Delta(r_1, r_2) = 8$ *and* $M \cong W(T_0; -1)$ *or* $W(T_0; 5)$.

The proofs of these three results follow from the intersection graph methods we discussed in Section 4.1, where we take $\widehat{P}_1 \subset M(T; r_1)$, $\widehat{P}_2 \subset M(T; r_2)$ to be essential spheres or tori depending on the theorem to be proved. Let P_1, P_2 be the associated punctured surfaces, and $\Gamma_1 \subset \widehat{P}_1$, $\Gamma_2 \subset \widehat{P}_2$ the associated graphs.

The most delicate arguments involve the proof of Theorem 4.14. Under the assumptions of this theorem, the weaker bound $\Delta(r_1, r_2) \leqslant 2$ was obtained by Wu in [106] using clever, though relatively simple, arguments based on Scharlemann cycles. In [50] it was assumed that an example realizing the distance 2 existed, and a contradiction was obtained through the use of a complex combinatorial analysis.

Perhaps the most interesting argument is found in the proof of Theorem 4.17. Upon the assumption that $\Delta(r_1, r_2) \geqslant 6$ it is shown that there are only four possible configurations for the triple $(P_1, P_2, P_1 \cap P_2)$ and further, that for each of these configurations, any regular neighbourhood of $P_1 \cup P_2 \cup T$ in an orientable 3-manifold has a boundary consisting of T and a collection of 2-spheres. The irreducibility of M then implies that there are only four possibilities for M, and that these are in fact the four exceptional manifolds listed in the theorem.

Table 2 summarizes the situation when M is assumed to be compact, connected, orientable, and not to contain any essential disks, spheres, annuli, or tori. Its entries are upper bounds for the distance between boundary slopes whose associated fillings yield surfaces of the specified types. All the bounds are sharp.

The entries not obtainable from the three theorems above are Disk–Disk (Theorem 4.3), Disk–Annulus [53], Disk–Torus [52], Sphere–Annulus [111], Annulus–Annulus and Annulus–Torus [45]. The case Disk–Sphere is due to M. Scharlemann and may be deduced from the following fundamental result.

Recall from Definition 4.2 the notion of a cable space.

DEFINITION 4.18. We say that the pair (M, T) is *cabled* if $M = C \cup_S M_1$ where $S \subset \operatorname{int}(M)$ is an essential torus and C is a cable space with boundary $T \cup S$.

If (M, T) is cabled, say $M = C \cup_S M_1$ where C is a cable space, then there is a slope on T corresponding to the essential annulus described in Example 4.1. It turns out that any cabling of (M, T) yields the same slope, and so we shall refer to it as the *cabling slope* of (M, T).

Table 2

	Disk	Sphere	Annulus	Torus
Disk	1	0	2	2
Sphere	–	1	2	3
Annulus	–	–	5	5
Torus	–	–	–	8

THEOREM 4.19 (Scharlemann [90]). *Let M be a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold and T be a toral component of ∂M. Suppose that there is a slope r_1 on T for which $M(T; r_1)$ is ∂-reducible. Then for any slope $r \neq r_1$ on T, one of the following four possibilities occurs.*

(i) $M(T; r_1) \cong M(T; r) \cong S^1 \times D^2$.

(ii) *(M, T) is cabled, r_1 is its cabling slope, $M(T; r_1) \cong S^1 \times D^2 \# L$ (L a lens space), and $M(T; r) \cong S^1 \times D^2$.*

(iii) *(M, T) is cabled and r is its cabling slope.*

(iv) *$M(T; r)$ is irreducible and no torus component of $\partial M \setminus T$ compresses in both $M(T; r_1)$ and $M(T; r)$.*

REMARK 4.20. Gabai [36] and Berge [6] have classified the knots in $S^1 \times D^2$ which admit non-trivial surgeries yielding $S^1 \times D^2$, and as a consequence determined exactly when parts (i) or (ii) of the theorem arise. In particular they showed that if part (i) of the theorem occurs then M is the exterior of a braid in $M(T; r_1) \cong S^1 \times D^2$.

The proof of Theorem 4.19 involves the foliation-free version of sutured manifold theory developed by Scharlemann in [89]. The main technical innovation introduced in [90] shows that if a sutured manifold admits an appropriate compressing disk or essential 2-sphere, then it has a sutured manifold hierarchy (see the proof of Theorem 4.10) where each term also has this property. The intersection graph method is then introduced under the assumption that $M(T; r)$ is either reducible or ∂-reducible, and in particular a special type of configuration called a *Gabai disk* is used to construct essential annuli $(A, \partial A) \subset (M, T)$. A regular neighbourhood N of $A \cup T$ is homeomorphic to the product of a twice-punctured disk with I, and if one of the components of $\partial N \setminus T$ compresses in M, then it is not hard to see that (M, T) is cabled with cabling slope the slope of the given annulus.

Observe that if we assume that $M(T; r_1)$ is ∂-reducible and $M(T; r)$ is reducible, then Scharlemann's theorem implies that (M, T) is cabled, say $M = C \cup_S M_1$. In particular, M admits an essential torus. This accounts for the "0" in the Disk–Sphere entry Table 2.

4.5. *Laminations and Dehn filling*

The notion of an essential surface in a 3-manifold has been generalized by Gabai and Oertel [38] to the notion of an essential lamination. We refer the reader to their paper for a more detailed explanation of the technical terms below.

An *essential lamination* is a foliation $\mathcal{L}$ of a closed subset of int(M) by smooth surfaces such that

- $\mathcal{L}$ contains no spherical leaf or toral leaf which bounds a solid torus;
- each closed complementary region of $\mathcal{L}$ is irreducible and has an incompressible, end-incompressible boundary.

There is also a notion of an *essential lamination with boundary*, though we will not go into the details here.

An essential surface is an example of an essential lamination, but there are many *laminar* manifolds, i.e., manifolds which contain an essential lamination, which do not contain any essential surfaces (see, e.g., Example 5.2 of Gabai and Oertel [38], or the discussion following Theorem 4.22 below). Nevertheless laminar manifolds are constrained topologically in that they are irreducible, ∂-irreducible, and their universal cover is $\mathbf{R}^3$ [38]. Thus it is of interest to ask when an essential lamination in M survives Dehn filling. The answer to this question is that generically they do, as Y.-Q. Wu showed in proving a version of Theorem 4.3 in this setting [107].

One instance where laminations arise naturally occurs when M is a surface bundle over the circle with pseudo-Anosov monodromy. The latter condition implies that the monodromy can be taken to preserve a 1-dimensional lamination of the surface whose complementary regions are either cusped disks or once-punctured cusped disks. The suspension of this lamination becomes an essential lamination in M whose complementary regions are either solid tori or solid tori with a central core removed.

THEOREM 4.21 (Gabai and Oertel [38]). *Let K be a fibred knot with pseudo-Anosov monodromy. Then there is a slope r_0 of K such that if $\Delta(r, r_0) > 1$, then $M_K(r)$ is laminar.*

PROOF. Let M be the exterior of K, S its fibre, $f : S \to S$ its pseudo-Anosov monodromy, and $\mathcal{L}$ the suspension of the stable lamination of f. Each complementary region of $\mathcal{L}$ may be identified with $U \times I/(x, 0) = (r(x), 1)$ where U is either

– a disk with $k \geqslant 2$ (outward) cusps if $U \cap \partial M = \emptyset$, or

– a once-punctured disk with $k \geqslant 1$ cusps otherwise,

and $r : U \to U$ is a rotation by some angle of the form $2\pi/m$. Let r_0 be the slope on ∂M parallel to the curves determined by the suspension of the cusps of U.

For each slope r on ∂M, we can consider $\mathcal{L}$ as a lamination in the manifold $M(r)$. By construction $\mathcal{L}$ has no leaves which are spheres or tori which bound a solid torus. Furthermore the precise knowledge we have of the structure of the complementary regions of $\mathcal{L} \subset M(r)$ implies readily that each such region is irreducible and has an incompressible, end-incompressible boundary as long as $\Delta(r, r_0) \geqslant 2$. Thus $\mathcal{L}$ remains essential in $M(r)$ when $\Delta(r, r_0) \geqslant 2$. □

Surgeries on other families of knots have also been successfully studied via laminations. C. Delman used the Hatcher–Thurston method of classification of essential surfaces in 2-bridge knot exteriors ([56]) to prove the following result.

THEOREM 4.22 (Delman [26]). *The exterior of a non-torus 2-bridge knot admits a lamination which remains essential after any non-trivial surgery.*

Delman employed a lamination version of Menasco's two annulus criterion (Theorem 4.8) to see that the laminations he constructed were *persistent*, i.e., remain essential after non-trivial surgery. It is interesting to contrast this theorem with the fact that since a 2-bridge knot exterior contains no closed, essential surface [56] and has only finitely many boundary slopes (Theorem 2.9), all but finitely many surgeries on such a knot are non-Haken.

The flexibility of essential laminations, as compared to essential surfaces, was aptly demonstrated by Naimi [77] and Roberts [84], who showed that in many instances, for each non-trivial slope r of a knot, there is an essential lamination with boundary consisting of simple closed curves of slope r (compare Theorem 2.9). Moreover, capping it off yields an essential lamination in the manifold obtained by surgery along that slope. Naimi gave another (independent) proof of Theorem 4.22 in this way, while Roberts studied surgery on certain alternating knots.

The interested reader is directed to D. Gabai's survey article [37] on the role of laminations in Dehn surgery theory.

4.6. *Heegaard splittings and Dehn filling*

We noted in Section 2.1 that the filling operation amounts to attaching a 2-handle/3-handle pair. Thus if Σ is a Heegaard surface for M (i.e., $\Sigma \subset \operatorname{int}(M)$ is a closed, connected, orientable surface which splits M into two submanifolds, each obtainable from Σ by 2 and 3-handle additions), then it is also a Heegaard surface for any filling of M. In particular if N is a compact, connected, orientable 3-manifold and we define the *Heegaard genus* of N to be $g(N) = \min\{\operatorname{genus}(\Sigma) \mid \Sigma \text{ is a Heegaard surface for } N\}$, then

$$g\big(M(T;r)\big) \leqslant g(M)$$

for all slopes r on T. Note that it is possible for the genus of $M(T;r)$ to drop significantly. For instance for any $n > 0$, it is relatively simple to construct a knot $K \subset S^3$ whose group $\pi_1(S^3 \setminus K)$ requires at least n generators, so $g(M_K) \geqslant n$. On the other hand, $g(M_K(\mu_K)) = g(S^3) = 0$. Moriah and Rubinstein [75] showed that a drop in genus after filling should be considered an exceptional event. The content of Theorem 4.23 makes this more concrete.

THEOREM 4.23 (Rieck [83]). *Let M be a compact, connected, orientable, irreducible 3-manifold and $T \subset \partial M$. If there are no essential annuli in the pair (M, T) and*

$$r_1, r_2 \in \big\{r \in \mathit{Slope}(T) \mid g(M(T;r)) \leqslant g(M) - 2\big\},$$

then $\Delta(r_1, r_2) < 36(g(M) - 2)^2 + 54(g(M) - 2) + 9$. In particular

$$g(M) - 1 \leqslant g\big(M(T;r)\big) \leqslant g(M)$$

for all but at most finitely many slopes r on T.

REMARK 4.24. The constants 36, 54, and 9 which appear in the statement of the theorem arise from the method of proof and undoubtedly do not give a sharp inequality.

PROOF. The proof is based on the intersection graph method that we have encountered several times already, though setting up the appropriate surfaces in this instance turns out to be non-trivial.

Let r_1, r_2 be two slopes on T for which there are Heegaard surfaces $\Sigma_1 \subset M(T;r_1)$, $\Sigma_2 \subset M(T;r_2)$ of genus no larger than $g(M) - 2$. Denote by K_1, K_2 the cores of the filling solid tori in $M(T;r_1)$, $M(T;r_2)$.

From the definition of a Heegaard surface, it follows that there are 1-complexes $X_j \subset M(T; r_j)$ whose complements can be identified with $\Sigma_j \times \mathbf{R}$ for $j = 1, 2$. We start by moving K_j off X_j and then isotoping the foliation $\Sigma_j \times \mathbf{R}$ so that K_j is in *thin position*. This means that the composition $h_j : K_j \to \Sigma_j \times \mathbf{R} \to \mathbf{R}$ is Morse and that the number of intersections of K_j with a certain collection of leaves of the foliation (comprised of one leaf lying between each successive pair of critical levels of h) is minimized. Without loss of generality we may assume that $h_j(K_j) = [0, 1]$.

Y. Rieck shows [83] that the inequality genus$(\Sigma_j) \leqslant g(M) - 2$ implies that K_j is not isotopic to a curve on Σ_j, and hence has non-trivial thin position. A delicate analysis of the intersections between the level sets $\Sigma_1 \times \{s\}$ and $\Sigma_2 \times \{t\}$ reveals that there are $t_1, t_2 \in (0, 1)$ such that if $P_1 = M \cap (\Sigma_1 \times \{t_1\})$ and $P_2 = M \cap (\Sigma_2 \times \{t_1\})$, then after a small isotopy of P_2, every arc of $P_1 \cap P_2$ is essential in both P_1 and P_2. If we assume that $\Delta(r_1, r_2) \geqslant 36(g(M) - 2)^2 + 54(g(M) - 2) + 9$, then by studying families of intersection arcs in $P_1 \cap P_2$, it can be shown that there are two such arcs which are parallel on both P_1 and P_2. The disks realizing these parallelisms glue together to form an essential annulus in (M, T), contrary to hypothesis. Thus $\Delta(r_1, r_2) < 36(g(M) - 2)^2 + 54(g(M) - 2) + 9$, as claimed. □

The theorem is nicely illustrated by the following example.

EXAMPLE 4.25. The exterior M of the right-handed trefoil knot K has Heegaard genus 2. Its fillings are described in Example 5.2. In particular it follows that if r_0 is the $+6$ slope on ∂M, then

$$g\big(M(T; r)\big) = \begin{cases} 0 & \text{if } r = \mu_K, \\ 1 & \text{if } \Delta(r, r_0) = 1 \text{ and } r \neq \mu_K, \\ 2 & \text{if } \Delta(r, r_0) > 1. \end{cases}$$

5. The topology of filled manifolds

Consider a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold M whose boundary contains a torus T. In this section we discuss the relationship between the topological and geometric properties of M with those of the fillings $M(T; r)$. Our point of departure is Thurston's work on 3-manifold topology which implies, for instance, that if ∂M consists of incompressible tori, then M satisfies one of the following three conditions (cf. Chapter 3):

(i) M is a Seifert fibred space;
(ii) M contains an essential torus;
(iii) M is hyperbolic.

This is a particularly fruitful way to consider M with regard to the study of its fillings, as the defining properties of each of the three categories suggest powerful methods of analysis. We shall see in the following three subsections that for a generic slope r on T, the manifolds $M(T; r)$ inherit the defining property of the category, and we shall describe the current state of knowledge concerning the non-generic slopes.

5.1. *Fillings of Seifert fibred spaces*

This is by far the simplest of the three cases. Indeed we shall see in Theorem 5.1 below that the fillings of such manifolds are readily classified. We begin with a brief introduction to Seifert fibred manifolds. For a more complete discussion see W. Jaco's monograph [60].

It follows from the work of Epstein [28] that a *fibred solid torus*, i.e., a solid torus C^∞ foliated by circles, is fibre-preserving diffeomorphic to one of the standard models described in the next paragraph.

Let p, q be a pair of coprime integers where $p \geqslant 1$, and let D^2 be the unit disk in $\mathbf{C}$. The *fibred solid torus of type* (p, q), denoted by $V_{(p,q)}$, is the quotient space $D^2 \times I/\{(x, 1) = (e^{2\pi i q/p}x, 0)\}$ endowed with the foliation by circles (called *fibres*) induced from the I-factors in $D^2 \times I$. Each $V_{(p,q)}$ is homeomorphic to $S^1 \times D^2$ and further $V_{(p,q)} \cong V_{(p',q')}$ by a fibre-preserving homeomorphism if and only if $p' = p$ and $q' \equiv \pm q \pmod p$.

We shall say that a fibred solid torus V is of type (p, q) if it is fibre-preserving homeomorphic to $V_{(p,q)}$. Both p and the $(\bmod\, p)$ class of $\pm q$ are invariants of V, p being called its *index*. Consideration of the standard models shows that if r_0, r_1, r_2 are the slopes on ∂V corresponding respectively to a fibre, a meridian, and a dual slope to a meridian, then

$$p = \Delta(r_0, r_1) \quad \text{and} \quad q \equiv \pm\Delta(r_0, r_2) \pmod p.$$

The sign of q in the congruence above can be determined by paying close attention to orientations.

A Seifert fibred structure on M is a C^∞ foliation of M by circles (called *fibres*) such that each fibre Φ has a closed tubular neighbourhood, consisting entirely of fibres, which is fibre-preserving homeomorphic to some fibred solid torus. The index $p \geqslant 1$ of this fibred solid torus is called the *index* of Φ. A fibre Φ of index p is referred to as an *exceptional* fibre if $p > 1$, and a *regular* fibre otherwise. For instance in a fibred solid torus of type (p, q), all fibres are regular except perhaps for the core fibre $\Phi_0 = (\{0\} \times I)/(0, 0) = (0, 1)$ which has index p. It follows that exceptional fibres are isolated and lie in the interior of M. Thus the exceptional fibres are finite in number, say $\Phi_1, \Phi_2, \ldots, \Phi_n$. Furthermore the foliation induces (by restriction) a locally-trivial S^1 bundle whose total space is $M_0 = M \setminus (\Phi_1 \cup \Phi_2 \cup \cdots \cup \Phi_n)$. Observe that ∂M is foliated by regular fibres and so it consists of a collection of tori.

The *orbit space* B of M is the space of leaves of the given foliation and topologically is a compact surface whose boundary consists of the fibres lying on ∂M. There is more structure inherent in the orbit space though. Namely, each point may be tagged with the index of the corresponding fibre in M. Let $x_i \in \operatorname{int}(B)$ correspond to the exceptional fibre Φ_i and let p_i be its tag. The Seifert structure on M then determines a 2-dimensional orbifold $\mathcal{B}$, called the *base orbifold* of M, whose underlying space is B and whose singular set is $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a cone point of index p_i. We say that $\mathcal{B}$ has the form $B(p_1, p_2, \ldots, p_n)$. For instance the base orbifold of a fibred solid torus of type (p, q) is of the form $D^2(p)$.

Once M is endowed with a Seifert structure, each of its boundary components has a distinguished *fibre slope* r_0 which corresponds to a fibre of the given Seifert structure. The key to understanding the topology of the manifolds $M(T; r)$ lies in observing that as long as $r \neq r_0$, the Seifert structure on M extends across the filling solid torus.

THEOREM 5.1 (Heil [58]). *Let M be a compact, connected Seifert fibred space with base orbifold $\mathcal{B}$ of the form $B(p_1, p_2, \dots, p_n)$, $T \subset \partial M$ a torus, C the component of ∂B corresponding to T, and r_0 the slope on T corresponding to a fibre of the Seifert fibration of M. If*

(i) *$r \neq r_0$, then the Seifert structure on M extends over $M(T; r)$ in such a way that its base orbifold is $\mathcal{B} \cup_{C=\partial D^2} D^2(\Delta(r, r_0))$. In fact the induced foliation on T extends over the filling solid torus so that it becomes a fibred solid torus of type $(\Delta(r, r_0), \Delta(r, r_1))$, where r_1 is any slope on T dual to r_0.*

(ii) *$r = r_0$, then*

$$M(T; r_0) \cong \left(\#_{i=1}^{n} L(p_i, q_i)\right) \# \left(\#_{j=1}^{b} S^1 \times D^2\right) \# \left(\#_{k=1}^{m} S^1 \times S^2\right),$$

where the ith exceptional fibre of M is of type (p_i, q_i), b is the number of boundary components of $M(T; r_0)$, and m is twice the genus of B when B is orientable, and the number of $\mathbf{R}P^2$-factors of B otherwise.

PROOF. Circle foliations on $S^1 \times S^1$ are determined up to isotopy by the slope of any leaf. Hence consideration of the standard fibred solid tori reveals that all circle foliations of $S^1 \times S^1$ extend across $S^1 \times D^2$ as long as the leaves are not meridians. Thus if V denotes the filling solid torus of $M(T; r)$, our hypothesis that $r \neq r_0$ implies that the foliation on T induced from the Seifert structure from M extends over V. As we observed above, the type (p, q) of V can be computed as $p = \Delta(r_0, r)$ and $q \equiv \pm\Delta(r, r_1) \pmod p$ where r_1 is any slope on T dual to r_0. Thus assertion (i) holds.

Assertion (ii) can be justified in two steps. First consider the case where $n = 0$. The surface B may be sliced into a 2-disk by cutting it open along an appropriate collection of disjoint arcs. Hence M can be cut open along the essential annuli lying above the given arcs so that the result is a solid torus M_0. These annuli intersect T in a collection of curves of slope r_0 and if the arcs are chosen carefully, the annuli may be capped off in $M(T; r_0)$ to form m spheres and b disks. These spheres and disks split the filling solid torus of $M(T; r_0)$ into a collection of 2-handles and the conclusion of (ii) is reached by seeing how the solid torus M_0, the annuli, and the 2-handles piece together to form $M(T; r_0)$.

Finally assume that $n > 0$ and apply the previous case to the exterior N, say, of the exceptional fibres of M. By paying close attention to how a fibred neighbourhood of one such fibre attaches to its $S^1 \times D^2$ factor of $N(T; r_0)$ it can be shown that each such attachment produces an $L(p_j, q_j)$ summand of $M(T; r_0)$. □

EXAMPLE 5.2 (Moser [76]). Think of the (m, n) torus knot K as lying on the standard, unknotted torus which splits S^3 into the union of two solid tori. From this point of view it is easy to see that its exterior M admits a Seifert structure which splits as the union of two solid fibred tori, one of type (m, n) and the other of type (n, m). Its base orbifold has the form $D^2(m, n)$. The fibre slope r_0 on ∂M corresponds to the integral fraction mn (cf. Proposition 2.4), and so by Theorem 5.1, $M_K(mn) \cong L(m, n) \# L(n, m)$. If $p/q \neq mn$, Theorem 5.1 shows that $M_K(p/q)$ admits a Seifert structure with base orbifold $S^2(|m|, |n|, |p - mnq|)$. In the interesting special case when $|p - mnq| = 1$ (i.e., the distance of the p/q slope to r_0 is 1), $M_K(p/q)$ is the lens space $L(p, n^2q)$.

5.2. *Fillings of manifolds which contain an essential torus*

In this section we shall assume that M contains an essential torus S. Set

$$Compr(S) = \{r \in Slope(T) \mid S \text{ compresses in } M(T; r)\}.$$

It is considered a good situation when a slope $r \notin Compr(S)$. For instance Theorem 4.14 implies that the fillings associated to all but at most three slopes $r \notin Compr(S)$ are Haken manifolds, and so are subject to the various structure theorems associated to such spaces. Also when $r \notin Compr(S)$ much useful information can be gleaned from the fact that the characteristic family of tori in $M(T; r)$ (cf. Chapter 3) usually coincides with the characteristic family of tori in M.

The set $Compr(S)$ is fairly well understood. For slopes $r \in Compr(S)$, the torus S compresses to a 2-sphere in $M(T; r)$. If this 2-sphere is inessential then S bounds a solid torus in $M(T; r)$. Indeed taking M_1 to be the closure of the component of $M \setminus S$ which contains T, then $M_1(T; r)$ is either reducible or $S^1 \times D^2$. In the latter case S must separate M, say $M = M_1 \cup_S M_0$, and so there is a slope r' on S for which $M(T; r) \cong M_0(S; r')$, i.e., the r-filling of M coincides with the r'-filling of the "simpler" manifold M_0. It is an immediate consequence of Scharlemann's theorem (Theorem 4.19) that this is the generic case.

THEOREM 5.3. *Suppose that* $\#Compr(S) > 1$. *Then S separates M, say $M = M_1 \cup_S M_0$ where $T \subset \partial M_1$, and for all $r \in Compr(S)$, but at most one, we have $M_1(T; r) \cong S^1 \times D^2$. Furthermore, if $M_1(T; r)$ is reducible for some $r \in Compr(S)$ then (M, T) is cabled and r is the cabling slope.*

PROOF. Let M_1 be the closure of the component of $M \setminus S$ which contains T. We observed above that for each $r \in Compr(S)$, $M_1(T; r)$ is either reducible or $S^1 \times D^2$. If the latter occurs then S separates M.

Suppose that $r_1, r_2 \in Compr(S)$ are distinct and $M_1(T; r_2)$ is reducible. Since $M_1(T; r_1)$ is ∂-reducible, Theorem 4.19 implies that (M_1, T) is cabled and r_2 is the cabling slope. Since the cabling slope of (M_1, T) is unique, i.e., independent of the choice of cabling of (M_1, T), a similar argument shows that $M_1(T; r_1)$ must be irreducible, and therefore homeomorphic to $S^1 \times D^2$. The theorem is now readily deduced. □

In the situation where there are at least two slopes $r \in Compr(S)$ for which $M(T; r) \cong S^1 \times D^2$, Gabai [36] and Berge [6] have determined both the possible topological types for M_1 and the set $Compr(S)$.

The following result provides a useful criterion for the existence of a cable space in M cobounded by S and T.

THEOREM 5.4 (Culler, Gordon, Luecke and Shalen [21]). *Let M be a compact, connected, orientable, irreducible 3-manifold and T a toral boundary component of M. If r_1 and r_2 are two slopes on T such that S compresses in $M(r_1)$ and $M(r_2)$, then either*

(i) $\Delta(r_1, r_2) \leqslant 1$, *or*

(ii) *T and S cobound a cable space $M_1 \subseteq M$ such that if r_0 is the cabling slope of (M, T), then S compresses in $M(T; r)$ if and only if $\Delta(r_0, r) \leqslant 1$.*

PROOF. Cut M open along S and replace M by M_1, the component of the resulting manifold which contains T. Then $M_1(T; r_1)$ and $M_1(T; r_2)$ are ∂-reducible and so if $\Delta(r_1, r_2) > 1$, Theorem 4.3 implies that there is an annulus A properly embedded in M_1 whose boundary consists of an essential curve on S', a parallel copy of S in M_1, and an essential curve $C_0 \subset T$. Furthermore if r_0 denotes the slope of C_0 on T, then S' compresses in $M_1(T; r)$ if and only if $\Delta(r_0, r) \leqslant 1$. We must see that M_1 is a cable space.

Let K be the core of the filling solid torus V of $M_1(T; r_1)$. Now as $\Delta(r_0, r_1) = 1$, the annulus $A \subset M_1$ may be extended into V to produce a new annulus whose boundary is $K \cup C_0$. In particular this implies that K is isotopic to a curve in S'. It follows that any 2-sphere in $M_1(T; r_1)$ may be isotoped off K, and therefore into the irreducible manifold M_1. Hence $M_1(T; r_1)$ is irreducible. But then since S' compresses in $M_1(T; r_1)$, $M_1(T; r_1)$ is a solid torus. Moreover combining the facts that K is isotopic to a curve on $\partial M_1(T; r_1)$, T is incompressible in M_1, and S' is not parallel in M_1 to T, we conclude that M_1 is a cable space. □

The discussion above suggests that the properties of $M(T, r)$ should be studied according to whether or not $r \in Compr(S)$ and whether or not (M, T) is cabled. It shows that if $\#Compr(S) > 1$, then S is separating, say $M = M_1 \cup_S M_0$ where $T \subset \partial M_1$, and for all but at most one slope $r \in Compr(S)$, there is a slope r' on S such that $M(T; r) = M_0(S; r')$. The most striking example of this phenomenon occurs when T and S cobound a cable space C, say $M = C \cup_S M_0$. If C has type (p, q) (cf. Example 4.1), then C is homeomorphic to the exterior of a regular fibre lying in the interior of a fibred solid torus of type (p, q). In particular C admits a Seifert fibring with base orbifold of the form $(S^1 \times I)(|q|)$, and for which r_0, the cabling slope of (M, T), is the fibre slope. Theorem 5.4 shows $Compr(S) = \{r \in Slope(T) \mid \Delta(r_0, r) \leqslant 1\}$ and in fact Theorem 5.1 gives $C(T; r_0) \cong (S^1 \times D^2) \# L(p, q)$ while $C(T; r) \cong S^1 \times D^2$ if $\Delta(r_0, r) = 1$. Now the inclusions $S, T \subset C$ induce an isomorphism $H_1(T; \mathbf{Q}) \to H_1(S; \mathbf{Q})$ and therefore there is a natural correspondence $r \leftrightarrow r'$ between the slopes on T and those on S (if $\Delta(r_0, r_1) = 1$ then r' is characterised by the condition that it be the meridional slope of the solid torus $C(T; r)$) such that

$$M(T; r) \cong \begin{cases} M_0(S; r_0') \# L(p, q) & \text{if } r = r_0, \\ M_0(S; r') & \text{if } \Delta(r, r_0) = 1. \end{cases}$$

It is useful to note that if r_1, r_2 are slopes on T satisfying $\Delta(r_1, r_0) = \Delta(r_2, r_0) = 1$, then Gordon [41] proved that

$$\Delta(r_1', r_0') = \Delta(r_2', r_0') = |q| \quad \text{and} \quad \Delta(r_1', r_2') = q^2 \Delta(r_1, r_2).$$

5.3. *Fillings of hyperbolic manifolds*

Consider a compact, connected, orientable, irreducible, ∂-irreducible 3-manifold M whose boundary contains a torus T. Let $\mathcal{T}$ be the union of all tori in ∂M. We shall say that M is *hyperbolic* if $M \setminus \mathcal{T}$ admits a complete, finite volume Riemannian metric of constant sectional curvature -1 such that $\partial M \setminus \mathcal{T}$ is totally geodesic. If M admits a hyperbolic structure, then this structure is unique up to isometry. Chapter 3 of this volume serves as a good reference on geometric structures.

W. Thurston's *hyperbolization theorem* states that a compact, orientable, irreducible, ∂-irreducible 3-manifold will be hyperbolic as long as it contains no essential tori or annuli, is non-Seifert (i.e., does not admit the structure of a Seifert fibred manifold), and is Haken. His *hyperbolization conjecture* contends that we may drop the requirement that the manifold be Haken from the previous assertion. The converse of the conjecture is known – a hyperbolic manifold contains no essential spheres, disks, tori or annuli and is non-Seifert. Our goal in this section is to discuss what is known about

$$\mathcal{E}(M;T) = \{r \in Slope(T) \mid M(T;r) \text{ is not hyperbolic}\},$$

the set of *exceptional* slopes on T.

Thurston [95] studied the metric completions of the set of incomplete hyperbolic structures on M which are close to the given complete one. In particular he saw that if we identify $Slope(T)$ with $\pm$-pairs of primitive elements in $\mathbf{Z} \oplus \mathbf{Z} \cong H_1(T)$, then for slopes $\pm(p,q)$ with $|p| + |q| \gg 0$, there is some such incomplete structure which extends to a complete hyperbolic structure on $M(T; \pm(p,q))$ (cf. Chapter 3). Consequently he deduced

THEOREM 5.5 (Thurston [95]). *Suppose that M is a compact, connected, orientable, hyperbolic 3-manifold. Then $\mathcal{E}(M;T)$ is finite.*

One of the main goals of research on Dehn filling has been to understand the finite set $\mathcal{E}(M;T)$, at least to the extent of determining universal bounds on

$$e(M;T) = \#\mathcal{E}(M;T),$$

universal bounds on the distance between (families of) slopes in $\mathcal{E}(M;T)$, and constraints on the hyperbolic manifolds M which admit exceptional surgeries. Empirical evidence suggests that $e(M;T)$ is small, that the slopes in $\mathcal{E}(M;T)$ are close to one another, and that manifolds admitting several exceptional slopes are also in some sense small.

When $\partial M \neq T$, Thurston's hyperbolization theorem combines with the results of Section 4.4 to produce an essentially optimal result.

THEOREM 5.6 (Gordon [45]). *Suppose that M is hyperbolic and $T \neq \partial M$ is a toral boundary component of M. If r_1, r_2 are two slopes on T such that $M(T;r_1)$, $M(T;r_2)$ are not hyperbolic, then $\Delta(r_1,r_2) \leqslant 5$.*

PROOF. Observe that as $\partial M \neq T$, a filling of M is either Haken or reducible. Hence according to Thurston's hyperbolization theorem, $\mathcal{E}(M;T)$ consists of those slopes r for which $M(T; r)$ is either reducible, or ∂-reducible, or toroidal, or annular, or admits a Seifert structure. From Theorems 4.14, 4.15, and 4.17 we see that the distance between any two slopes on T that produce reducible or toroidal fillings is at most 5. On the other hand if $M(T; r)$ is annular, ∂-reducible or Seifert it admits either an essential disk or an essential annulus. We can now appeal to Table 2 to complete the proof. □

Notice that Theorem 5.6 implies that when $\partial M \neq T$, $e(M; T) \leqslant 8$ (see Table 1), but it is unknown whether or not $e(M; T) > 6$ is possible. The next two examples realize $e(M; T) = 6$, and the second shows that the distance bound 5 in Theorem 5.6 is sharp.

EXAMPLE 5.7 (*The Whitehead link*). Recall the Whitehead link L (cf. Example 4.16). Its exterior W admits an involution which exchanges its boundary components, and so the spaces resulting from filling along a component T of ∂W are independent of the choice of T. Thurston [95] showed that W is hyperbolic and Neumann and Reid [78] have calculated $\mathcal{E}(W; T) = \{1/0, 0, 1, 2, 3, 4\}$. Thus $e(W; T) = 6$ and $\max\{\Delta(r_1, r_2) \mid r_1, r_2 \in \mathcal{E}(W; T)\} = 4$.

EXAMPLE 5.8 (*The Whitehead sister link*). The *Whitehead sister link* L' is the $(-2, 3, 8)$ pretzel link pictured in Figure 8. Let T be the boundary component of W' corresponding to the unknotted component of L'. J. Weeks and C. Hodgson have studied this example using the computer programme *Snappea* (see the final paragraph of this section) and shown that $\mathcal{E}(W; T) = \{1/0, 2, 3, 7/2, 11/3, 4\}$. Thus $e(W; T) = 6$ and $\max\{\Delta(r_1, r_2) \mid r_1, r_2 \in \mathcal{E}(W; T)\} = 5$.

For the remainder of this section we shall concentrate on the case where ∂M is a torus T. The added complication that now arises is that the fillings of M will not, in general, be Haken manifolds, and so we cannot invoke Thurston's hyperbolization theorem. (Though if the first Betti number of M is at least 3, one may argue, as above, that the conclusion of Theorem 5.6 holds.) Indeed we must now cope with the possibility that we have fillings which either have finite fundamental groups or are closed, non-Haken Seifert fibred manifolds. The complexity thus introduced is such that currently there is no result remotely comparable to Theorem 5.6 available. C. Hodgson and S. Kerchoff have announced that there is some calculable constant $N \gg 0$ such that for any hyperbolic manifold M with toral boundary, $e(M) \leqslant N$, though this result should be tempered with the following example which constitutes the worst known case of manifold M in terms of the size of both $e(M) = e(M; T)$ and $\Delta(M) = \max\{\Delta(r_1, r_2) \mid r_1, r_2 \in \mathcal{E}(M; T)\}$.

EXAMPLE 5.9. Let K be figure 8 knot and M its exterior. In [95] it is shown that M is hyperbolic and $\mathcal{E}(M) = \{1/0, 0, \pm 1, \pm 2, \pm 3, \pm 4\}$. Thus $e(M) = 10$ and $\Delta(M) = 8$.

Gordon [43] gives an exhaustive discussion of the known manifolds M with relatively large collections of exceptional slopes and it is interesting to note that they all are fillings of the Whitehead link exterior. Consideration of these examples leads to the following conjecture.

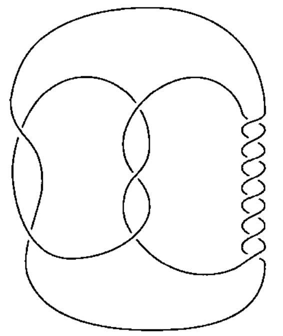


Fig. 8. The Whitehead sister link.

CONJECTURE 5.10. If M is a compact, connected, orientable, hyperbolic 3-manifold whose boundary is a torus then $e(M) \leqslant 10$ and $\Delta(M) \leqslant 8$. Moreover if W is the Whitehead link exterior, T a component of ∂W, and $M \not\cong W(T; -1)$, $W(T; 5)$, $W(T; 5/2)$, or $W(T; -2)$, then $\Delta(M) \leqslant 5$ and $e(M) \leqslant 8$.

One way to bridge the gap between examples and results is to work modulo the hyperbolization conjecture, an approach which has yielded results which are quite sharp. We explain this now. Define seven subsets of $\mathcal{E}(M; T)$ as follows. For $r \in Slope(T)$ let

- $r \in \mathrm{RED}(M)$ if and only if $M(r)$ is reducible;
- $r \in \mathrm{CYC}(M)$ if and only if $\pi_1(M(r))$ is cyclic;
- $r \in \mathrm{FIN}(M)$ if and only if $\pi_1(M(r))$ is finite;
- $r \in \mathrm{SFS}(M)$ if and only if $M(r)$ admits the structure of a Seifert fibred space;
- $r \in \mathrm{ASFS}(M)$ if and only if $r \in \mathrm{SFS}(M)$, $\pi_1(M(r))$ is infinite, and $M(r)$ is irreducible and contains no essential tori;
- $r \in \mathrm{TSFS}(M)$ if and only if $M(r)$ admits the structure of an irreducible Seifert fibred space which contains an essential torus; and
- $r \in \mathrm{TOR}(M)$ if and only if $M(r)$ admits an essential torus.

Set

$$\mathcal{E}_{TOP}(M) = \mathrm{RED}(M) \cup \mathrm{FIN}(M) \cup \mathrm{TOR}(M) \cup \mathrm{ASFS}(M) \subseteq \mathcal{E}(M).$$

The hyperbolization conjecture contends that $\mathcal{E}_{TOP}(M)$ coincides with $\mathcal{E}(M)$. Take, for instance, M to be figure 8 knot exterior (cf. Example 5.9). Then $\mathcal{E}(M) = \mathcal{E}_{TOP}(M)$ and $M_K(1/0) = S^3 \in \mathrm{CYC}(M)$, $M_K(0)$, $M_K(\pm 4) \in \mathrm{TOR}(M)$, and $M_K(\pm 1)$, $M_K(\pm 2)$, $M_K(\pm 3) \in \mathrm{ASFS}(M)$.

THEOREM 5.11 (Bleiler and Hodgson [8]). *Suppose that M is compact, connected, orientable, hyperbolic and ∂M is a torus. Then there are at most* 24 *slopes in $\mathcal{E}_{TOP}(M)$, and the distance between any two such slopes is at most* 22.[1]

[1] *Added in proof.* Agol [3] and Lackenby [66] have greatly sharpened the estimates in this theorem. They show that the lower bound of 2π referred to in its proof can be lowered to 6. Subsequently it can argued that there are at most 12 slopes in $\mathcal{E}_{TOP}(M)$ and the distance between any two such slopes is at most 10.

Table 3

	RED	CYC	FIN	ASFS	TSFS	TOR
RED	1#	1#	5*	6*	3	3#
CYC	–	1#	2#	10	1#	10
FIN	–	–	3#	10	5*	10
ASFS	–	–	–	10	10	10
TSFS	–	–	–	–	5	5
TOR	–	–	–	–	–	8#

PROOF. The proof of this theorem involves a careful interpretation of work of M. Gromov and W. Thurston. By hypothesis int(M) admits a complete hyperbolic metric of finite volume. A *cusp* of M is a closed subset of int(M) homeomorphic to $S^1 \times S^1 \times [0, \infty)$ such that $S^1 \times S^1 \times \{0\}$ lifts to a disjoint collection of horoballs in hyperbolic 3-space, the universal cover int(M). Fix a cusp C of M and let $M \setminus \text{int}(C) = M_0 \cong M$. After a suitable normalization, work of Gromov and Thurston shows that if r is a slope on ∂M_0 such that any representative curve has length at least 2π, then the restriction of the given hyperbolic metric to M_0 extends to a Riemannian metric on $M_0(r)$ all of whose sectional curvatures are strictly negative. It is known that such a manifold is irreducible, contains no essential tori, is not Seifert fibred, and has infinite fundamental group. Therefore the associated slope is not an element of $\mathcal{E}_{TOP}(M)$. The theorem is then proved by combining an analysis of the possibilities for the set of slopes having a representative curve of length less than 2π together with the estimates for the cusp volume of M (the supremum of the volumes of its cusps) due to Adams [1]. □

Adams [2] further refined this result through the use of the cusp volume of M. He showed that as cusp volume increases, $\#\mathcal{E}_{TOP}(M)$ decreases, eventually dropping to 1 when the cusp volume is larger than some fixed constant. He also examined the relation between the Heegaard genus of M and cusp volume to derive topological constraints on M from the existence of exceptional slopes.

The bounds obtained in Theorem 5.11 are very much larger than the conjectured ones. By restricting our attention to various families of slopes which lie in $\mathcal{E}_{TOP}(M)$ we can do much better. This is borne out in the following table whose entries give an upper bound for the distance between two exceptional slopes of the indicated types. In the table a sharp designates a sharp result, that is an upper bound realized by an example. An asterisk indicates that there are some exceptions (Table 3) to the result.

The following exceptions hold.

* Δ(RED, FIN): Assume the reducible filling does not yield $\mathbf{P}^3 \# \mathbf{P}^3$, the connected sum of two copies of projective 3-space.
* Δ(RED, ASFS): Assume the reducible filling is not $\mathbf{P}^3 \# \mathbf{P}^3$ or $S^1 \times S^2$.
* Δ(FIN, TSFS): Assume the TSFS filling is not a union of two twisted I-bundles over the Klein bottle.

The justifications for the bounds described in the table are due to I. Agol, S. Boyer, M. Culler, C.McA. Gordon, M. Lackenby, J. Luecke, S. Oh, P. Shalen, Y.-Q. Wu, and X. Zhang. Six of the bounds, namely the RED-RED, RED-TOR, RED-TSFS, TSFS-TSFS,

TSFS-TOR, and TOR-TOR entries, follow from Theorems 4.14, 4.15, and 4.17. The entries equalling 10 are derived in [3] and [66], while the remaining ones are consequences of the results discussed in the following section (see Theorems 5.12, 5.13, 5.15, 5.18).

A good feel for the geometry of fillings can be developed using the computer programme SnapPea written by J. Weeks. It is available from the University of Minnesota's Geometry Center via anonymous ftp at ftp://geom.umn.edu/pub/software/snappea. The input is a link diagram and slopes for some of the link's components. The algorithm then determines whether the resulting filled manifold admits a hyperbolic structure, and if so, it calculates many of the manifold's geometric invariants. For example its fundamental group, volume, group of isometries, and length spectrum.

5.4. *Small fillings*

In this section we discuss fillings which yield spaces which are in one way or another small. The term "small" has been used in several different senses in the literature, though in the most common one a 3-manifold is called *small* if it contains no closed essential surfaces. We begin with a discussion of fillings which yield manifolds having "small" fundamental groups. The following result is known as the *cyclic surgery theorem.*

THEOREM 5.12 (Culler, Gordon, Luecke and Shalen [21]). *Let M be a compact, connected, orientable, irreducible 3-manifold whose boundary is an incompressible torus. Assume further that M is not a Seifert fibred manifold. If r_1, r_2 are two slopes on ∂M such that $\pi_1(M(r_1))$ and $\pi_1(M(r_2))$ are cyclic, then $\Delta(r_1, r_2) \leqslant 1$.*

Example 5.2 shows that each torus knot admits infinitely many different surgeries which yield lens spaces, and therefore the restriction to non-Seifert manifolds in the cyclic surgery theorem is necessary. On the other hand, applying it to a non-torus knot $K \subset S^3$ (the exterior of such a knot is not Seifert fibred) shows that besides the meridional slope μ_K, any cyclic filling slope is integral. Further, there are at most two such non-meridional slopes, and if two, they correspond to successive integers. An example is provided by the $(-2, 3, 7)$ pretzel knot. Its exterior M was studied by Fintushel and Stern [31] who showed that it is hyperbolic and $M(18) \cong L(18, 5)$, $M(19) \cong L(19, 7)$.

PROOF OF THEOREM 5.12. The cyclic surgery theorem is proven through a combination of the intersection graph techniques of C.McA. Gordon and J. Luecke (Section 4.1) with the character variety methods pioneered by M. Culler and P. Shalen in [23,22]. The proof splits into the four cases (i) M is hyperbolic and neither r_1 nor r_2 is a strict boundary slope (a *strict* boundary slope is a slope corresponding to the boundary of an essential surface which is not the fibre of any fibration of M over the circle); (ii) the first Betti number of M is at least 2; (iii) the first Betti number of M is 1 and one of r_1 and r_2 is a strict boundary slope; (iv) the first Betti number of M is 1 and M contains an essential torus. Case (i) is dealt with through the use of the methods of Culler and Shalen while the proofs of the remaining cases use Gordon and Luecke's techniques. Suppose that r_1, r_2 are two slopes on ∂M such that $\pi_1(M(r_1))$ and $\pi_1(M(r_2))$ are cyclic.

Case (i). The set of characters of representations of $\pi_1(M)$ in $SL_2(\mathbf{C})$ forms a complex affine set $X(M)$ of dimension at least 1. Culler and Shalen show that to each curve $X_0 \subset X(M)$ can be associated a seminorm $\|\cdot\|_{X_0} : H_1(\partial M; \mathbf{R}) \to [0, \infty)$ such that for each $\alpha \in H_1(\partial M)$, $\|\alpha\|_{X_0}$ is the degree of the regular function $f_\alpha : X_0 \to \mathbf{C}$ given by $f_\alpha(\chi) = \chi(\alpha)^2 - 4$ (we are identifying α with its image, well-defined up to conjugation, under $H_1(\partial M) \equiv \pi_1(\partial M) \subset \pi_1(M)$). For $j = 1, 2$, denote by $\alpha(r_j)$ a primitive homology classes carried by r_j (cf. Section 2.4). A key fact proved in [21] is that if X_0 contains the character of an irreducible representation, then the fact that r_j is not a strict boundary slope combines with the cyclicity of $\pi_1(M(r_j))$ to give

$$\|\alpha(r_j)\|_{X_0} \leqslant \|\alpha\|_{X_0} \quad \text{for all } \alpha \in H_1(\partial M) \text{ such that } \|\alpha\|_{X_0} \neq 0.$$

In our situation there is a canonical choice of a curve X_0, namely one which contains the character of a discrete, faithful representation provided by the hyperbolic structure on M. It turns out that in this case $\|\cdot\|_{X_0}$ is a norm and the minimality of $\|\alpha(r_1)\|_{X_0}$ and $\|\alpha(r_2)\|_{X_0}$ translates into the bound $\Delta(r_1, r_2) \leqslant 1$. See Chapter 19 for a detailed account of the character variety method and applications to surgery theory.

Case (ii). In this case the first Betti number of any filling of M is at least 1. Thus $\pi_1(M(r_1)) \cong \pi_1(M(r_2)) \cong \mathbf{Z}$. Hence for $j = 1, 2$, any non-separating surface in $M(r_j)$ compresses to an essential 2-sphere $\widehat{P}_j$. Intersection graphs can then be formed (Section 4.1) and since neither $M(r_1)$ nor $M(r_2)$ has a lens space summand, the graphs have no Scharlemann cycles (Proposition 4.7). The same graph theoretic arguments employed in the proof of Theorem 4.3 now show that $\Delta(r_1, r_2) \leqslant 1$.

Case (iii). Theorem 4.13 implies that there is a closed essential surface in M which remains incompressible in $M(r')$ as long as $\Delta(r_1, r') > 1$. It follows that $\Delta(r_1, r_2) \leqslant 1$.

Case (iv). Since $b_1(M) = 1$, any torus in int(M) separates and so we can find one, S say, for which $M = M_1 \cup_S M_0$ where $\partial M_1 = (\partial M) \cup S$, $\partial M_0 = S$, and M_0 contains no essential tori. According to Thurston's work, M_0 is either hyperbolic or Seifert fibred.

We show that the hypothesis $\Delta(r_1, r_2) > 1$ leads to a contradiction. Under this assumption, Theorem 5.4 shows that M_1 is a cable space. Let r'_0 be the slope on S corresponding to a Seifert structure on M_1. The discussion at the end of Section 5.2 shows that there are slopes r'_1, r'_2 on S and an integer $n \geqslant 2$ such that for $j = 1, 2$, $M(r_j) = M_0(r'_j)$ and further

$$\Delta(r'_1, r'_2) = n^2 \Delta(r_1, r_2) \geqslant 8 \quad \text{while } n = \Delta(r'_1, r'_0) = \Delta(r'_1, r'_0). \tag{$*$}$$

Thus by the case (iii) of the proof, neither r'_1 nor r'_2 can be a strict boundary slope, and so by case (i) M_0 is not a hyperbolic manifold. From the classification of *atoroidal* Seifert manifolds (i.e., those which contain no essential tori) it can be shown that M_0 admits a Seifert structure with base orbifold of the form $D^2(p, q)$, $2 \leqslant p, q$. If ϕ is the slope on S corresponding to a fibre of this structure then $r'_0 \neq \phi$ as M admits no Seifert structures, and so $\Delta(r'_0, \phi) > 0$. We finally observe that the fact that $M_0(r'_1)$ and $M_0(r'_2)$ have cyclic fundamental groups forces $1 = \Delta(r'_1, \phi) = \Delta(r'_2, \phi)$ (see Theorem 5.1). It is now elementary to combine the latter relations with the identities $(*)$ above to see that $\Delta(r_1, r_2) = 1$ (cf. the argument on p. 137 of [46]), contrary to our hypotheses. This completes the argument. □

Next we consider the *finite surgery theorem.*

THEOREM 5.13 (Boyer and Zhang [13,15]). *Let M be a compact, connected, orientable, irreducible 3-manifold whose boundary is an incompressible torus. Assume further that M is neither a Seifert fibred manifold nor a cable on the twisted I-bundle over the Klein bottle.*

(i) *There are no more than six slopes r on ∂M such that $\pi_1(M(r))$ is either finite or infinite cyclic. Further, for any two such slopes r_1, r_2 we have $\Delta(r_1, r_2) \leqslant 5$.*
(ii) *If M is hyperbolic, then there are no more than five slopes r on ∂M such that $\pi_1(M(r))$ is either finite or infinite cyclic. Further, for any two such slopes r_1, r_2 we have $\Delta(r_1, r_2) \leqslant 3$.*
(iii) *If M is hyperbolic and r_1, r_2 are two slopes such that $\pi_1(M(r_1))$ is cyclic and $\pi_1(M(r_2))$ is finite, then $\Delta(r_1, r_2) \leqslant 2$.*

REMARK 5.14. Each of the three parts of this theorem is sharp. It is shown in [8] that the bounds 6 and 5 of part (i) are realized by taking M to be the exterior of the $(11, 2)$ cable of the trefoil knot, while the bounds 5 and 3 of part (ii) are realized when M is the exterior of figure 8 sister knot [105]. The same example shows that part (iii) is sharp. Note that this part of the theorem implies that if M is the exterior of a hyperbolic knot in the 3-sphere, then $M(p/q)$ has an infinite fundamental group whenever $|q| \geqslant 3$.

PROOF. The proof begins along the same general lines as that of Theorem 5.12 and results from [21] can be used to reduce to the case where M is hyperbolic and r is not a strict boundary slope. In this case, Boyer and Zhang show that if $\pi_1(M(r))$ is finite and X_0 is any curve in the $SL_2(\mathbf{C})$-character variety of M, then though $\|\alpha(r)\|_{X_0}$ may not be minimal, it almost is. In fact the difference between $\|\alpha(r)\|_{X_0}$ and the minimal norm of a non-zero element of $H_1(\partial M)$ can be computed in terms of the set of characters in X_0 of representations of the finite group $\pi_1(M(r))$. In the case where X_0 is the canonical curve, the bounds on the norms of homology classes associated to slopes whose fillings have finite fundamental groups suffice to prove part (iii) of the theorem [13], and also that part (ii) holds for all but finitely many possibilities for $\|\cdot\|_{X_0}$ [13,14]. To deal with the remaining cases, the relationship between the Culler–Shalen norm $\|\cdot\|_{X_0}$ and the A-polynomial associated to X_0 [21] is analysed, and then exploited. (See Chapter 19 for more details.) Finally, part (i) of the theorem is a consequence of part (ii) [13]. □

In an interesting application of these techniques, Tanguay [93] has shown that if M is the exterior of a hyperbolic 2-bridge knot, then $\pi_1(M(p/q))$ is infinite for all non-trivial slopes p/q (compare Theorem 4.22).

Another instance where the seminorm method can be used to determine good bounds on the distance between a slope r_0 and slopes $r \in \mathrm{FIN}(M)$ occurs when there are curves $X_0 \subset X(M(r_0)) \subset X(M)$ [16]. When $r_0 \in \mathrm{RED}(M)$ or $\mathrm{TSFS}(M)$ such curves can often be found and since the associated Culler–Shalen seminorms are indefinite (i.e., $\|\alpha(r_0)\|_{X_0} = 0$) it can be shown that there is a non-negative integer s such that $\|\alpha(r')\|_{X_0} = s\Delta(r', r_0)$ for any slope r' on ∂M. Thus when $s \neq 0$, $\|\cdot\|_{X_0}$ is equivalent to $\Delta(\cdot, r_0)$. We discussed above the fact that if $r_1 \in \mathrm{FIN}(M)$, then the seminorm of a class $\alpha(r_1)$ is usually small, and therefore $\Delta(r_1, r_0)$ is also. An example of the results they obtained is the following.

A Seifert fibred space is called a *small-Seifert* if it admits a Seifert structure whose orbifold is the 2-sphere with 3 or fewer cone points.

THEOREM 5.15 (Boyer and Zhang [16]). *Let M be a compact, connected, orientable, 3-manifold whose boundary is a torus which is neither Seifert fibred nor a union of a Seifert fibred manifold and a cable space. Assume that $M(r_1)$ is either reducible or a Seifert fibred space which is not small-Seifert.*

(i) *If $M(r_2)$ has a cyclic fundamental group then $\Delta(r_1, r_2) \leqslant 1$.*
(ii) *If $M(r_2)$ has a finite fundamental group and $M(r_1) \neq \mathbf{R}P^3 \# \mathbf{R}P^3$ or the union of two copies of the twisted I-bundle over the Klein bottle, then $\Delta(r_1, r_2) \leqslant 5$.*

Next we consider fillings which yield small manifolds with infinite fundamental groups. We suppose that M is a compact, connected, orientable, hyperbolic 3-manifold whose boundary is a torus.

The least understood family of exceptional slopes on ∂M is the set ASFS(M) whose fillings yield irreducible, atoroidal Seifert manifolds (see Table 3). In fact these fillings are all small-Seifert. We will discuss two of the approaches which have been developed in order to understand them.

An essential lamination (Section 4.2) is called *genuine* if it has a complementary region which is not an I-bundle. Thus genuine laminations do not extend to foliations of the ambient manifold. Our interest in them arises from the fact that Brittenham [18] has shown that small-Seifert spaces do not admit them. This gives us a way to analyze the set of small-Seifert filling slopes. As an example Brittenham observed that this method could be used to enhance Theorem 4.21 in the following way.

THEOREM 5.16 (Brittenham [18]). *Let K be a fibred knot with pseudo-Anosov monodromy. Then there is a slope r_0 of K such that if $\Delta(r, r_0) > 2$, then $M_K(r)$ is not a small-Seifert manifold.*

In another example Brittenham and Wu [19] studied the essential laminations in the exterior of hyperbolic 2-bridge knots constructed by Delman [26] and determined they could often produce ones which were genuine. Delman (see Theorem 4.22) had already shown that his laminations usually stay essential in all non-trivial surgeries, thus providing effective obstructions to a filling being small-Seifert. This led them to a precise determination of $\mathcal{E}_{TOP}(M)$ when M is the exterior of a hyperbolic 2-bridge knot.

THEOREM 5.17 (Brittenham and Wu [19]). *Let K be the p/q 2-bridge knot and suppose that $r \in \mathcal{E}_{TOP}(M_K)$, $r \neq \mu_K$. Then there are integers b_1, b_2 such that $|b_j| > 1$ and $p/q = 1/(b_1 - 1/b_2)$. Furthermore if $|b_1|$, $|b_2| > 2$ then there is a unique such r and it lies in TOR. Otherwise K is a twist knot and the set $r \in \mathcal{E}_{TOP}(M_K)$ can be explicitly determined.*

Another approach to the study of the slopes in ASFS(M) based on the characteristic submanifold theory of W. Jaco, P. Shalen, and K. Johannson has been employed by S. Boyer,

M. Culler, P. Shalen, and X. Zhang. They obtain bounds on the distance between reducible or toroidal filling slopes and small-Seifert filling slopes. An example of this work is as follows.

THEOREM 5.18 (Boyer, Culler, Shalen and Zhang [10]). *Suppose that M is a compact, connected, orientable, hyperbolic 3-manifold whose boundary is a torus and that r_1, r_2 are slopes on ∂M such that $M(r_1)$ is reducible and $M(r_2)$ is small-Seifert. If $M(r_1)$ is neither $S^1 \times S^2$ nor $\mathbf{R}P^3 \# \mathbf{R}P^3$, then $\Delta(r_1, r_2) \leqslant 6$.*

PROOF. The key feature of a small-Seifert space which is invoked in the proof is that either its fundamental group does not contain a non-abelian free group or it contains an essential, immersed torus.

The hypotheses imply that there is an essential planar surface F in M with non-empty boundary of slope r_1. Since M is hyperbolic, F has $n \geqslant 3$ boundary components. Now if $\pi_1(M(r_2))$ does not contain a non-abelian free group then there is a homotopically non-trivial knot in $M \setminus F$ which bounds a singular disk $f : D^2 \to M(r_2)$. Otherwise we can find an essential immersion $f : T^2 \to M(r_2)$. In either case if V denotes the r_2 filling solid torus, then the map f may be chosen so that $f^{-1}(F \cup V)$ is a graph consisting of (fat) vertices and essential curves. A standard argument (see, e.g., Lemma 4.1 of [46]) implies that if $\Delta(r_1, r_2) \geqslant 7$ then it is possible to find a set of $n + 1$ parallel edges in this graph. Each successive pair of these arcs gives rise to an essential homotopy of a wedge of two circles starting and ending in F and which passes through $M \setminus F$. W. Jaco, P. Shalen and independently, K. Johannson, proved that such homotopies have to pass through canonical I-bundles in M. The fact that there are n successive such homotopies can be used to show that parts of the I-bundles may be pieced together to build an essential torus in M, contrary to its hyperbolicity. □

Another class of small manifolds consists of those that are not virtually Haken (a manifold is *virtually Haken* if it is finitely covered by a Haken manifold). Set

$$\mathcal{VH}(M) = \{r \in Slope(\partial M) \mid M(r) \text{ is virtually Haken}\}.$$

It was conjectured by Waldhausen [102] that any compact, connected, orientable, irreducible 3-manifold with infinite fundamental group is virtually Haken, and so $\mathcal{VH}(M)$ is expected to coincide with $\mathrm{RED}(M) \cup \mathrm{FIN}(M)$.

Quite a lot can be said about the non-virtually Haken filling slopes when M is a Seifert fibred manifold or contains an incompressible, non-boundary parallel torus (see Hempel's paper [59]), and if M is hyperbolic, the truth of the Waldhausen conjecture would imply, via Theorems 4.14 and 5.13, that $\mathcal{VH}(M)$ contains no more than eight slopes. At present even the finiteness of $\mathcal{VH}(M)$ is unknown. For instance combined work of M. Baker, J. Hempel, S. Kojima and D. Long, and A. Nicas has shown that roughly 70% of the fillings of the exterior of figure 8 knot are virtually Haken manifolds, though the status of the remaining fillings is still open (cf. [5] for a more thorough discussion of this example, and problem 3.2 of R. Kirby's problem list [64] for the general problem). In the papers of Cooper and Long [20] and Boyer and Zhang [17] it is shown how certain hypotheses on a slope r_0

can be used to prove that $M(r)$ is virtually Haken as long as $\Delta(r, r_0)$ is large enough. For example Cooper and Long prove that m/n surgery on a non-fibred genus g knot in S^3 is virtually Haken as long as $|m| = \Delta(m/n, 0/1) \geqslant 12g - 4$, while Boyer and Zhang show that if $M(r_0)$ is reducible, but neither $S^1 \times S^2$ nor $\mathbf{R}P^3 \# \mathbf{R}P^3$, then $r \in \mathcal{VH}(M)$ as long as $\Delta(r, r_0) \geqslant 6$.

6. Surgery on knots in the 3-sphere

Many of the most interesting questions about surgery relate to exceptional surgeries on knots in the 3-sphere. Here we examine when non-trivial surgery on a knot $K \subset S^3$ can yield either S^3, a simply-connected manifold, a reducible manifold, or (when K is a hyperbolic knot) a non-hyperbolic manifold.

We saw in Section 5 that it is fruitful to invoke Thurston's view of 3-manifold topology when studying fillings. This is particularly true in the study of surgery on knots in S^3 where its application leads to a partition of these knots into three mutually disjoint families:
- the torus knots are those knots whose exteriors admit a Seifert fibring;
- the satellite knots are those knots whose exteriors admit an essential torus;
- the remaining knots are those whose exteriors are hyperbolic.

6.1. *Surgery and the knot complement problem*

Tietze [100] asked if the complement of a knot in the 3-sphere determines its knot type. In other words if two knots $K, K' \subset S^3$ have homeomorphic complements, or equivalently [27], homeomorphic exteriors, is there a self-homeomorphism of S^3 taking K to K'? This is the *knot complement problem*. For instance the trivial knot is determined by its complement because any knot whose exterior is a solid torus has a 2-disk for a Seifert surface, and so is unknotted. C.McA. Gordon and J. Luecke obtained a solution of the general problem using surgery theory.

THEOREM 6.1 (Gordon and Luecke [48]). *Knots in the 3-sphere are determined by their complements.*

PROOF. Let K be a knot in the 3-sphere which, from the remarks above, we may assume is knotted. Our first step is to reformulate the problem in terms of surgery.

Suppose that K and K' represent distinct knot types and that there is a homeomorphism $f : M_K \to M_{K'}$ between their exteriors. Note that the image of μ_K by f cannot be $\mu_{K'}$, for otherwise the restriction of f to $\partial N(K)$ could be extended to a homeomorphism $N(K) \to N(K')$, thereby producing a homeomorphism of S^3 taking K to K'. Let $r \neq \mu_K$ be the slope of K which is sent by f to $\mu_{K'}$. Then $M_K(r) \cong M_{K'}(\mu_{K'}) = S^3$. Hence K admits a nontrivial slope r such that $M_K(r) \cong S^3$. It is this fact that is shown to be impossible.

One of the key points, due independently to D. Gabai and C.McA. Gordon–J. Luecke, is to exploit Gabai's ideas on thin position. Denote by $\{\pm\infty\}$ the north and south poles of S^3 and observe that there is a natural identification of $S^3 \setminus \{\pm\infty\}$ with $S^2 \times \mathbf{R}$. We may assume that $K \subset S^2 \times \mathbf{R}$ and is in thin position with respect to the projection $S^2 \times \mathbf{R} \to \mathbf{R}$ (cf. the proof of Theorem 4.23). Also arrange the core of the filling solid torus to lie in thin position in $M_K(r) \cong S^3$. A careful analysis of the intersections between the two families of planar level sets in M_K coming from the 2-sphere foliations of $M_K(\mu_K) \setminus \{\pm\infty\}$ and $M_K(r) \setminus \{\pm\infty\}$ reveals that there are levels $P_1 \subset M_K$ from one family and $P_2 \subset M_K$ from the other, such that after a small isotopy of P_2, every arc of $P_1 \cap P_2$ is essential in both P_1 and P_2. Gordon and Luecke now play the two intersection graphs off each other. Using subtle combinatorial arguments, they show that either Γ_2 contains a Scharlemann cycle or Γ_1 satisfies a certain combinatorial hypothesis which implies that $H_1(M_K(r))$ contains non-trivial torsion. The impossibility of the latter implies the former must hold. But this in turn is impossible because if it did hold, Proposition 4.7 shows that $M_K(\mu_K) = S^3$ would have a lens space summand. This completes the argument. □

The knot complement problem can be posed for knots in an arbitrary closed, connected, orientable 3-manifold W: if two knots $K, K' \subset W$ have homeomorphic exteriors, is there a self-homeomorphism of W taking K to K'? The corresponding surgery problem asks if K is a knot in W and $M_K(r) \cong W$, does there exist a homeomorphism of M_K which takes μ_K to r? In this generality the answer is no. For instance Mathieu [72] found simple counterexamples by looking at surgeries on torus knots, though in each case, as with all subsequent examples, the homeomorphisms reversed orientation.

CONJECTURE 6.2 (*The generalized knot complement problem*). Suppose that K is a knot in a closed, orientable 3-manifold W and that $M_K(r) \cong W$ via an orientation preserving homeomorphism for some slope r of K. Then there is an orientation preserving homeomorphism of the exterior of K which takes μ_K to r.

Boileau, Domergue, and Mathieu [9] have shown that if K is a null-homotopic loop in W, then the conjecture holds for K as long as either $b_1(M_K) \geqslant 2$ or M_K contains an essential torus which does not cobound a cable space of type $(2, q)$ with ∂M_K. The following theorem shows that the conjecture holds for generic knots in $\mathbf{Z}$-homology 3-spheres.

THEOREM 6.3 (Boyer and Lines [11]). *Let K be a knot in a closed, connected, oriented 3-manifold W such that $H_1(W) = 0$. If $\Delta_K''(1) \neq 0$, then Conjecture 6.2 holds for K. In fact there is no orientation preserving homeomorphism between $M_K(r)$ and $M_K(r')$ if $r \neq r'$.*

PROOF. As $H_1(W) = 0$, there is a canonical identification of the slopes of K with $P^1(\mathbf{Q})$ (Theorem 2.4) such that $H_1(M_K(p/q)) \cong \mathbf{Z}/p$. Suppose then that there is an orientation preserving homeomorphism between $M_K(p/q)$ and $M_K(p'/q')$. For homological reasons we may suppose that $p = p' \geqslant 1$. It was shown by Boyer and Lines in [11] that if $\lambda(W)$ is the Casson invariant of W (see [4]) and $\sigma(K; p) = \sum_{j=1}^{|p|-1} \sigma_K(e^{2\pi i j/p})$ is the sum of the p-signatures of K [101], then

$$\bar{\lambda}\big(M_K(p/q)\big) = \lambda(W) + (q/2p)\Delta_K''(1) + (1/8|p|)\sigma(K; p) \in \mathbf{Q}$$

is an invariant of the oriented manifold $M_K(p/q)$. Hence as $\bar{\lambda}(M_K(p/q)) = \bar{\lambda}(M_K(p/q'))$ and $\Delta_K''(1) \neq 0$, we have $q = q'$, which was to be proved. □

For a more detailed discussion of this topic see problems 1.80 and 1.81 of Kirby's problem list [64].

6.2. *The property P conjecture*

The Poincaré conjecture contends that any closed, simply-connected 3-manifold is homeomorphic to the 3-sphere. As Theorem 3.1 implies that any counterexample to it could be represented by surgery on some link in S^3, it is natural to ask whether or not a counterexample can be obtained by surgery on a knot. We saw in the proof of Theorem 6.1 that no non-trivial surgery on a non-trivial knot in S^3 yields S^3 and so we are led to ask if there is a non-trivial knot $K \subset S^3$ and a non-trivial slope r of K for which $M_K(r)$ is simply-connected. The *property P* (for "Poincaré") *conjecture*, due independently to Gonzàlez-Acuña [39] and Bing and Martin [7], states that there is no such knot.

CONJECTURE 6.4 (*The property P conjecture*). If K is a non-trivial knot in the 3-sphere and r is a non-meridional slope of K, then $M_K(r)$ is not simply-connected.

We note that for any knot $K \subset S^3$ and slope r corresponding to the fraction p/q, $p \geqslant 0$, $H_1(M_K(p/q)) \cong \mathbf{Z}/p$, and so if $M_K(r)$ is simply-connected then $p = 1$.

We begin our analysis with the easiest case.

PROPOSITION 6.5. *Non-trivial torus knots satisfy the property P conjecture.*

PROOF. Suppose that K is the (m, n) torus knot and that $q \neq 0$ is an integer. According to Example 5.2, $M_K(1/q)$ is a Seifert fibred space with base orbifold of the form $S^2(|m|, |n|, |qmn - 1|)$. The quotient of $\pi_1(M_K(1/q))$ by the normal subgroup generated the homotopy class of a regular fibre is determined by this orbifold, and in this case it is the $(|m|, |n|, |qmn - 1|)$ triangle group (see [60, VI.9]). If K is assumed to be non-trivial we have $|m|, |n| \geqslant 2$, and so this triangle group is non-trivial. Hence $M_K(1/q)$ is not simply-connected. □

If we recall that the exteriors of satellite knots and hyperbolic knots are not Seifert fibred and that μ_K is a cyclic filling slope, then the following lemma follows from an application of the cyclic surgery theorem (Theorem 5.12).

LEMMA 6.6. *If K is either a satellite knot or a hyperbolic knot and r is a non-trivial slope for which $M_K(r)$ is simply-connected, then r is either the $+1$ or -1 slope.*

THEOREM 6.7 (Gabai [35]). *Satellite knots satisfy property P.*

PROOF. Suppose that K is a satellite knot and that $S \subset M_K$ is an essential torus. As S compresses in S^3, the 3-dimensional Schoenflies theorem implies that there is a solid torus $V \subset S^3$ whose boundary is S. Since S is essential in M_K it follows that (i) $K \subset \text{int}(V)$,

(ii) W, the exterior of K in V, is ∂-irreducible, and (iii) the core of V is a non-trivial knot K_0 which is not isotopic to K.

Suppose that r is a slope for which $M_K(r)$ is simply-connected. According to Lemma 6.6, r is an integral slope of the form $\varepsilon \in \{\pm 1\}$. Set $T = \partial M_K = \partial W \setminus S$ and observe that $M_K(r) = M_{K_0} \cup_S W(T; r)$. The non-triviality of K_0 implies that S must compress in $W(T; r)$, and so $W(T; r) \cong (S^1 \times D^2) \# M_1$ for some closed 3-manifold M_1. Hence if r_0 is the slope on S corresponding to a meridian of the $S^1 \times D^2$ factor of $W(T; r)$, then $M_K(r) \cong K_0(r_0) \# M_1$. In particular both $K_0(r_0)$ and M_1 are simply-connected. Now homological calculations show that if the winding number of K in V is ω, then the fraction associated to the slope r_0 is ε/ω^2. Thus using Proposition 6.5 and Lemma 6.6, the non-triviality of K_0 implies that $\omega \in \{-1, 0, 1\}$. On the other hand the only one of the four possibilities of Theorem 4.19 applied to $W(T; r)$ (with $r_1 = \mu_K$) which is consistent with our situation is the first: $W(T; r_1) \cong W(T; r) \cong S^1 \times D^2$. Then by Remark 4.20, K is a braid in V. But a braid in V of winding number $\omega \in \{-1, 0, 1\}$ is isotopic to the core of V, and this possibility is precluded by the fact that S is not ∂-parallel. Thus no non-trivial surgery on K yields a simply-connected manifold. □

Many families of hyperbolic knots are known to satisfy the property P conjecture, though it still remains an open problem. These families include symmetric knots [21], non-flat alternating knots [84], and arborescent knots [109]. C. Delman and R. Roberts have announced that the theory of essential laminations can be used to show that all hyperbolic alternating knots do as well.

We close this section with a lovely application of the Casson invariant. It states that the conjecture holds for generic knots.

THEOREM 6.8 (Casson). *If* $\Delta''_K(1) \neq 0$, *then* K *satisfies the property P conjecture.*

PROOF. The Casson invariant $\lambda(\Sigma)$, of a $\mathbf{Z}$-homology 3-sphere Σ, is an algebraic count of the number of characters of irreducible representations $\pi_1(\Sigma)) \to SU(2)$. In particular if $\lambda(\Sigma) \neq 0$, there is at least one such character, so $\pi_1(\Sigma) \neq \{1\}$. Now the Dehn surgery formula for this invariant (see [4]) shows that $\lambda(M_K(1/q)) = q\Delta''_K(1)$, which is non-zero when $q \neq 0$. Thus the knot K satisfies the property P conjecture. □

6.3. *The cabling conjecture*

Another well-known open problem is the *cabling conjecture*, due to Gonzàlez-Acuña and Short [40]. It addresses the question of when a surgery on a knot in the 3-sphere can yield a reducible manifold. This phenomenon does arise, as the following construction indicates.

Let q be an integer satisfying $|q| \geqslant 2$. A (p, q) cable of a knot $K_0 \subset S^3$ is a knot K which lies on $\partial N(K_0)$ and represents the p/q slope of K_0. The exterior of such a K decomposes as $M_K = C \cup M_{K_0}$ where C is a cable space of type (p, q). From Example 4.1, there is a slope r_1 on ∂M_K such that $C(\partial M_K; r_1) \cong S^1 \times D^2 \# L(q, p)$. Homological calculations show that r_1 is the pq slope on ∂M_K and it can be shown that $M_K(pq) \cong M_{K_0}(p/q) \# L(q, p)$. Hence as long as $M_{K_0}(p/q) \ncong S^3$, $M_K(pq)$ is reducible.

CONJECTURE 6.9 (*The cabling conjecture*). If K is a knot in the 3-sphere and r is a slope of K for which $M_K(r)$ is a reducible manifold, then K is a (p,q)-cable on a knot K_0, and r is the slope corresponding to the integer pq.

PROPOSITION 6.10. *Torus knots satisfy the cabling conjecture*

PROOF. First observe that if K is the (m,n) torus knot, then it is an (m,n) cable on the trivial knot. By Example 5.2, if $p/q \neq mn$, then $M_K(p/q)$ is a Seifert fibred space whose base orbifold is of the form $S^2(|m|, |n|, |p - qmn|)$. Such manifolds are known to be irreducible (cf. [60, Lemma VI.7]). On the other hand, $M_K(mn) \cong L(m,n) \# L(n,m)$. □

C.McA. Gordon and J. Luecke have found strong constraints on the form of $M_K(r)$ when it is reducible.

THEOREM 6.11 (Gordon–Luecke). *If $M_K(r)$ is reducible, then r is an integral slope and $M_K(r)$ has a lens space summand. Further, if M_K contains no closed, essential surface, then $M_K(r)$ is a connected sum of two lens spaces.*

PROOF. By Proposition 6.10, we may assume that K is not a torus knot. Hence as μ_K is a cyclic slope, Theorem 5.15 implies that the distance between r and μ_K is at most 1, thus r is an integral slope. Next we proceed as in the proof of Theorem 6.1, replacing the thin 2-sphere $\widehat{P}_2 \subset M_K(r)$ used there by an appropriately chosen essential 2-sphere. The same methods produce a Scharlemann cycle in Γ_2, and hence a lens space summand of $M_K(r)$.

Finally observe that as $M_K(r)$ has a lens space summand, it is not homeomorphic to $S^1 \times S^2$. We may therefore apply Theorem 4.13 to see that if M_K contains no closed, essential surface, then $M_K(r)$ is a connected sum of two lens spaces. □

Theorem 6.11 implies the following important result of D. Gabai (the line of the original argument can be found in the proof of Theorem 4.11).

COROLLARY 6.12 (Gabai [34]). *If K is a non-trivial knot in S^3, then $M_K(0)$ is irreducible.*

THEOREM 6.13 (Scharlemann [90]). *Satellite knots satisfy the cabling conjecture.*

PROOF. Let K be a satellite knot. We proceed in a manner similar to the proof of Theorem 6.7. Since K is a satellite knot, there is a solid torus $V \subset S^3$ whose boundary is an essential torus $S \subset M_K$ such that (i) $K \subset \text{int}(V)$, (ii) W, the exterior of K in V, is ∂-irreducible, and (iii) the core of V is a non-trivial knot K_0 which is not isotopic to K.

Suppose that $M_K(r) = M_{K_0} \cup W(T;r)$ is reducible. It is easy to see that $W(T;r)$ is either reducible or ∂-reducible and irreducible. In the former case, Theorem 4.19 implies that M_K is cabled and r is the cabling slope. Thus the theorem holds. In the latter $W(T;r) \cong S^1 \times D^2$ as it is a ∂-reducible, irreducible manifold bounded by a torus. Thus $M_K(r) = W(T;r) \cup M_{K_0} = M_{K_0}(r')$ for some slope r' of K_0. According to Remark 4.20, K is a braid in V of winding number ω, and our hypotheses show $|\omega| \geqslant 2$. Thus letting r

correspond to the fraction p (Proposition 6.11 implies that r is integral), then homological considerations show that r' corresponds to $p/\omega^2 q$ where $\omega^2 \geqslant 4$. But this contradicts Theorem 6.11. Thus the latter case does not arise and the theorem is proven. □

What about hyperbolic knots? The cabling conjecture predicts that all surgeries on hyperbolic knots yield irreducible manifolds. This is still an open problem, though the conjecture has been shown to hold for many families including genus 1 knots [12], alternating knots [74], arborescent knots [109], and symmetric knots [57].

6.4. *Hyperbolic surgery on knots in the 3-sphere*

We shall use the notation from Section 5.3. Let K be a hyperbolic knot in S^3 with exterior M_K. What can we say about the fractions corresponding to the slopes in $\mathcal{E}(M_K)$, and what can we say about K if $\mathcal{E}(M_K)$ contains more than just μ_K? These are the subjects we address here.

CONJECTURE 6.14 (*Hyperbolic surgery conjecture*). If K is a hyperbolic knot in the 3-sphere and $r = p/q$ is a slope of K for which $M_K(r)$ is not a hyperbolic manifold, then $|q| \leqslant 2$, and if $|q| = 2$ then $M_K(r)$ admits an essential torus and is not a Seifert fibred space.

Much is known about this conjecture, as Table 4 indicates.

Justifications for the first five of these bounds are to be found (from left to right) in [3, 13,21,47] and [16,66] (see Theorems 5.12, 5.13, and 5.15). The last one is a consequence of the next theorem which is proven using the intersection graph method.

THEOREM 6.15 (Gordon and Luecke [50,51]). *Let K be a hyperbolic knot in the 3-sphere and suppose that $M_K(p/q)$ contains an essential torus. Then $|q| \leqslant 2$ and if $|q| = 2$ then K is a strongly invertible knot whose exterior has Heegaard genus no larger than 2.*

Observe that this theorem places strong topological constraints on a hyperbolic knot satisfying the theorem's hypotheses, and it is expected that knots K for which $\mathcal{E}(M_K)$ contains a non-trivial slope are also subject to strong topological constraints. Another example of this phenomenon is described by C. Adams in [2], where he finds restrictions on the Heegaard genus of M_K. In the constructions given below, we present three ways to produce examples of exceptional surgeries on knots. It is interesting to note that in each case, all known exceptional surgeries of the given type arise from that construction.

Table 4

$M_K(p/q)$	RED	CYC	FIN	ASFS	TSFS	TOR
Upper bound for $\lvert q \rvert$	1	1	2	10	1	2

EXAMPLE 6.16 (*Eudave-Muñoz* [29]). The goal here is to construct hyperbolic knots for which some half-integral surgery contains an essential torus. First of all observe that if a knot J is the product of two non-trivial 2-string tangles, then the inverse image of the tangle 2-sphere in $\Sigma_2(J)$, the 2-fold cover of S^3 branched over J, is an essential torus. If we assume further that J has unknotting number 1 (Section 6.5), then $\Sigma_2(J)$ is of the form $M_K(p/2)$ for some knot $K \subset S^3$ and odd integer p (Proposition 6.19). In fact Theorem 6.15 can be used to show that any knot for which some $p/2$ surgery contains an essential torus is of unknotting number 1 and admits a non-trivial tangle decomposition. In [29] a large family of such knots is constructed and it has been suggested that this family contains all knots for which some half-integral surgery is a toroidal manifold.

EXAMPLE 6.17 (*Berge*). In an unpublished manuscript entitled *Some knots with surgeries yielding lens spaces*, J. Berge detailed the following elegant construction of knots in S^3 which admit surgeries yielding lens spaces.

Let $S^3 = H_1 \cup H_2$ be the standard genus 2 Heegaard splitting of the 3-sphere and consider a non-trivial knot $K \subset F = H_1 \cap H_2$. For $j = 1, 2$ K represents (up to conjugacy and taking inverse) an element of $\pi_1(H_j)$, a free group of rank 2. We shall say that K is *primitive* in H_j if this element is part of a basis of $\pi_1(H_j)$. If K is primitive and V is the manifold which results from attaching a 2-handle to H_j along K, then $\pi_1(V) \cong \mathbf{Z}$ and in fact $V \cong S^1 \times D^2$.

Now suppose that K is primitive in both H_1 and H_2 and let r_0 be the (integral) slope of K determined by one of its parallels lying on the surface F. Split the filling solid torus of $M_K(r_0)$ into two 3-cells A, B which intersect in a disjoint pair of meridian disks, and consider $M_K(r_0)$ as the union $(H_1 \cup A) \cup (H_2 \cup B)$. Now by choice of r_0, the attachment of A to H_1 and B to H_2 are just 2-handle attachments along K, and therefore by choice of K, $M_K(r_0)$ is a union of two solid tori. Since K is non-trivial and $r \neq \mu_K$, $M_K(r_0) \neq S^3$, $S^1 \times S^2$ (Theorem 6.1, Corollary 6.12), and therefore $M_K(r_0)$ is a non-trivial lens space.

Berge has given a complete list of these knots. It is conjectured that any surgery on a knot which yields a lens space arises from this construction.

There is a related construction, due to J. Dean, which produces surgeries yielding small-Seifert fibred spaces.

EXAMPLE 6.18 (*Dean* [24]). We continue with the notation of the previous example. In this case we look for non-trivial knots $K \subset F = H_1 \cap H_2$ which are primitive/Seifert, that is it is primitive in H_1 and (m, n) *Seifert fibred* in H_2. The latter means that the manifold obtained by attaching a 2-handle to H_2 along K admits a Seifert fibring with base orbifold of the form $D^2(m, n)$. For instance if the quotient of $\pi_1(H_2)$ by the normal subgroup generated by K is isomorphic to the group presented by $\langle x, y \mid x^m y^n \rangle$, then Dean shows that K is (m, n) Seifert fibred. Arguing as in the last example it can be seen that when $K \subset F$ is a primitive/Seifert knot and r_0 is the slope of K determined by F, then $M_K(r_0)$ is either a connected sum of two lens spaces or a Seifert fibred manifold whose base orbifold is the 2-sphere with 3 or fewer cone points. It is conjectured that any surgery on a knot which yields a Seifert fibred space arises from this construction. Note that this would imply that the only Seifert manifolds which can occur from surgery on a knot in the 3-sphere are small-Seifert manifolds.

6.5. *Surgery and unknotting numbers*

We finish this chapter with an application of surgery theory to the study of unknotting numbers.

A *crossing change* on a knot K involves isolating two disjoint arcs on K and passing one through the other as indicated in Figure 9. Any knot K may be transformed into a trivial knot by a finite sequence of crossing changes. The *unknotting number* of K, denoted by $u(K)$, is the minimal number of such changes needed. The following theorem describes a useful connection between unknotting number 1 knots and surgery theory.

PROPOSITION 6.19 (Lickorish [69]). *If $u(K) = 1$, then there are a knot $J \subset S^3$ and an odd integer p such that the 2-fold cyclic cover of S^3 branched over K is homeomorphic to $M_J(p/2)$.*

PROOF. Let B_0 be a small ball in which a crossing change transforming K to the unknot K_1 takes place (see Figure 9). Let $B_\infty = S^3 \setminus \text{int}(B_0)$. Let M_0, M_1, M_∞ be the 2-fold cyclic covers of B_0, B_1, B_∞ branched over $B_0 \cap K$, $B_0 \cap K_1$, $B_\infty \cap K$, respectively. If $\Sigma_2(K)$, $\Sigma_2(K_1)$ denote denote the 2-fold cyclic covers of S^3 branched over K and K_1, then $S^3 = \Sigma_2(K_1) = M_1 \cup M_\infty$. Now M_1 is a actually a solid torus, so M_∞ is the exterior of some knot $J \subset S^3$. Furthermore, as M_0 is also a solid torus and $\Sigma_2(K) = M_0 \cup M_\infty$, then $\Sigma_2(K) \cong M_J(r)$ for some slope r of J. Let p/q correspond to the slope r and suppose that $p \geqslant 0$. Then $p = |H_1(M_J(r))| = |H_1(\Sigma_2(K))| = |\Delta_K(-1)| \equiv 1 \pmod 2$. Finally a careful analysis of the covering map $M_0 \to B_0$ shows that $q = \pm 2$. □

Recall that a knot is called *prime* if it is non-trivial and each time it is expressed as a knot product $K = K_1 \# K_2$, one of K_1, K_2 is the trivial knot.

THEOREM 6.20 (Scharlemann [88]). *Unknotting number 1 knots are prime.*

PROOF. We use an argument originally due to C.McA. Gordon (see also [112]).

Let K be an unknotting number 1 knot. According to Proposition 6.19, there is a knot $J \subset S^3$ such that $\Sigma_2(K) \cong M_J(p/2)$ for some odd integer p. Hence by Theorem 5.15, $\Sigma_2(K)$ is irreducible.

On the other hand we claim that the 2-fold branched cover of a composite prime knot is reducible. To see this suppose that K' is a composite knot and fix a 2-sphere S in S^3

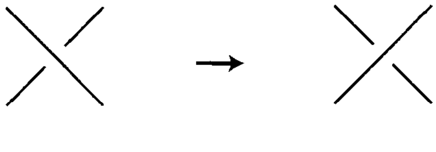


Fig. 9. A crossing change.

which intersects K' in exactly two points giving a product decomposition $K' = K_1 \#_S K_2$. Let B_1, B_2 be the balls in S^3 which bound S, indexed so that $A_j = K \cap B_j$ is the arc of K' made up from K_j.

The inverse image of S in $\Sigma_2(K')$ is a separating 2-sphere $\widetilde{S}$ which divides $\Sigma_2(K')$ into two pieces $\widehat{B}_1 \cup_{\widetilde{S}} \widehat{B}_2$ where $\widehat{B}_j$ is the 2-fold cover of B_j branched over A_j. It follows from F. Waldhausen's proof of the $\mathbf{Z}/2$ Smith conjecture [103] that as K_j is a non-trivial knot, $\widehat{B}_j$ is not a 3-ball. Thus $\Sigma_2(K')$ is reducible.

It follows that K must be a prime knot. □

The 2-bridge knots are characterized by the fact that their 2-fold branched cyclic covers are lens spaces. We denote by $K_{(p,q)}$ $(p > 1)$ the 2-bridge knot for which $\Sigma_2(K_{(p,q)}) \cong L(p,q)$.

As an example consider the knot $K_{(17,4)}$ (8_3 in the Alexander–Briggs notation) pictured in Figure 10. While this knot "clearly" has unknotting number 2, this was only proven through the use of the cyclic surgery theorem (Theorem 5.12).

THEOREM 6.21 (Kanenobu and Murakami [62]). *If a 2-bridge knot $K_{(p,q)}$ has unknotting number 1, then there exist relatively prime integers a and b with $2ab = p \pm 1$ and $K_{(p,q)}$ is equivalent to $K_{(p,2b^2)}$.*

PROOF. Let $K = K_{(p,q)}$. Now if $u(K) = 1$, then Proposition 6.19 implies that $L(p,q) \cong \Sigma_2(K)$ is of the form $M_J(n/\pm 2)$ for some knot $J \subset S^3$ and odd integer $n > 0$. Homological considerations show that we have $n = p$.

According to Theorem 5.12, the knot J must be the (a,b) torus knot for some relatively prime pair of integers a and b. According to Example 5.2, $M_J(p/2) \cong L(p,q)$ has base orbifold $S^2(|a|, |b|, |p - 2ab|)$. Hence the orbifold fundamental group of $S^2(|a|, |b|, |p - 2ab|)$, being a quotient of $\pi_1(L(p,q))$ [60, VI.9], is cyclic. But this is the $(|a|, |b|, |p - 2ab|)$ triangle group and hence as $|a|, |b| \geqslant 2$, $|p - 2ab| = 1$. Setting $\varepsilon = 2ab - p \in \{\pm 1\}$ we have $p = 2ab + \varepsilon$ and so $p/2 = (2ab + \varepsilon)/2$. Finally it is shown in [76] that $J((2ab+\varepsilon)/2) = L(2ab + \varepsilon, 2b^2) = L(p, 2b^2)$. □

As an example, suppose that $u(8_3) = u(K_{(17,4)}) = 1$. As $(17 \pm 1)/2$ equals $8 = 2^3$ or $9 = 3^2$, 8_3 is equivalent to $K_{(17,2)} = K_{(17,2\cdot 8^2)}$ or $K_{(17,2\cdot 9^2)} = K_{(17,9)}$. Hence by passing to branched covers we see that $L(17,4)$ is homeomorphic to $L(17,2)$ or $L(17,9)$. But this contradicts the classification of lens spaces: $L(p,q) \cong L(p,q')$ if and only if $q' \equiv \pm q^{\pm 1}$ (mod p). Thus $u(8_3) > 1$ and is indeed 2.

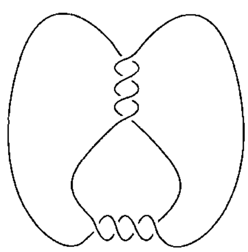


Fig. 10. The knot 8_3.

References

[1] C. Adams, *The noncompact hyperbolic 3-manifold of minimal volume*, Proc. Amer. Math. Soc. **100** (1987), 601–606.

[2] C. Adams, *Unknotting tunnels in hyperbolic 3-manifolds*, Math. Ann. **302** (1995), 177–195.

[3] I. Agol, *Topology of hyperbolic 3-manifolds*, Ph.D. thesis, U. C. San Diego (1998).

[4] S. Akbulut and J. McCarthy, *Casson's Invariant for Oriented Homology 3-Spheres*, Math. Notes, Vol. 36, Princeton University Press (1990).

[5] M. Baker, *On coverings of figure eight knot surgeries*, Pacific J. Math. **150** (1991), 215–228.

[6] J. Berge, *The knots in $D^2 \times S^1$ which have nontrivial Dehn surgeries that yield $D^2 \times S^1$*, Topology Appl. **38** (1991), 1–19.

[7] R.H. Bing and J.M. Martin, *Cubes with knotted holes*, Trans. Amer. Math. Soc. **155** (1971), 217–231.

[8] S. Bleiler and C. Hodgson, *Spherical space forms and Dehn fillings*, Topology **35** (1996), 809–833.

[9] M. Boileau, M. Domergue and Y. Mathieu, *Sur le complément d'un noeud homotope à zéro*, Preprint.

[10] S. Boyer, M. Culler, P. Shalen and X. Zhang, In preparation.

[11] S. Boyer and D. Lines, *Surgery formulae for Casson's invariant and extensions to homology lens spaces*, J. Reine Angew. Math. **405** (1990), 181–220.

[12] S. Boyer and X. Zhang, *Reducible Dehn surgery and toral Dehn surgery*, Topology Appl. **68** (1995), 285–303.

[13] S. Boyer and X. Zhang, *Finite Dehn surgery on knots*, J. Amer. Math. Soc. **9** (1996), 1005–1050.

[14] S. Boyer and X. Zhang, *A note on finite Dehn fillings*, Can. Math. Bull., to appear.

[15] S. Boyer and X. Zhang, *A proof of the finite filling conjecture*, Preprint.

[16] S. Boyer and X. Zhang, *On Culler–Shalen seminorms and Dehn filling*, Ann. of Math., to appear.

[17] S. Boyer and X. Zhang, *On virtually Haken Dehn filling*, Topology, to appear.

[18] M. Brittenham, *Essential laminations, exceptional Seifert-fibered spaces, and Dehn filling*, Preprint.

[19] M. Brittenham and Y.-Q. Wu, *The classification of Dehn surgeries on 2-bridge knots*, Preprint.

[20] D. Cooper and D. Long, *Virtually Haken Dehn filling*, Preprint.

[21] M. Culler, C. Gordon, J. Luecke and P. Shalen, *Dehn surgery on knots*, Ann. of Math. **125** (1987), 237–300.

[22] M. Culler and P. Shalen, *Bounded, separating surfaces in knot manifolds*, Invent. Math. **75** (1984), 537–545.

[23] M. Culler and P. Shalen, *Varieties of group representations and splitting of 3-manifolds*, Ann. of Math. **117** (1987), 109–146.

[24] J. Dean, Ph.D. thesis, University of Texas at Austin (1996).

[25] M. Dehn, *Über die Topologie des dreidimensionalen Raumes*, Math. Ann. **69** (1910), 137–168.

[26] C. Delman, *Essential laminations and Dehn surgery on 2-bridge knots*, Topology Appl. **63** (1995), 201–221.

[27] C.H. Edwards, *Concentricity in 3-manifolds*, Trans. Amer. Math. Soc. **113** (1964), 406–423.

[28] D.B.A. Epstein, *Periodic flows on three-manifolds*, Ann. of Math. **95** (1972), 66–82.

[29] M. Eudave-Muñoz, *Non-hyperbolic manifolds obtained by Dehn surgery on hyperbolic knots*, Geometric Topology, Vol. 1, W. Kazez, ed., AMS/IP Studies in Advanced Mathematics (1996), 35–61.

[30] R. Fenn and C. Rourke, *On Kirby's calculus of links*, Topology **18** (1979), 1–15.

[31] R. Fintushel and R. Stern, *Constructing lens spaces by surgery on knots*, Math. Z. **175** (1980), 33–51.

[32] D. Gabai, *Foliations and the topology of 3-manifolds*, J. Differential Geom. **18** (1983), 445–503.

[33] D. Gabai, *Foliations and the topology of 3-manifolds, II*, J. Differential Geom. **26** (1987), 461–478.

[34] D. Gabai, *Foliations and the topology of 3-manifolds, III*, J. Differential Geom. **26** (1987), 479–536.

[35] D. Gabai, *Surgery on knots in solid tori*, Topology **28** (1989), 1–6.

[36] D. Gabai, *1-bridge braids in solid tori*, Topology Appl. **37** (1990), 221–235.

[37] D. Gabai, *Problems in foliations and laminations*, Geometric Topology, Vol. 2, W. Kazez, ed., AMS/IP Studies in Advanced Mathematics (1996), 1–34.

[38] D. Gabai and U. Oertel, *Essential laminations in 3-manifolds*, Ann. of Math. **130** (1989), 41–73.

[39] F. González-Acuña, *Dehn's construction on knots*, Bol. Soc. Mat. Mex. **15** (1970), 58–79.

[40] F. González-Acuña and H. Short, *Knot surgery and primeness*, Math. Proc. Camb. Phil. Soc. **99** (1986), 89–102.

[41] C. Gordon, *Dehn surgery and satellite knots*, Trans. Amer. Math. Soc. **275** (1983), 687–708.

[42] C. Gordon, *Dehn surgery on knots*, Proceedings of the International Congress of Mathematicians, Kyoto, 1990, Springer, Tokyo (1991), 631–642.

[43] C. Gordon, *Dehn filling: A survey*, Proceedings of the Mini Semester in Knot Theory, Banach Center, Warsaw, Poland (1995).

[44] C. Gordon, *Combinatorial methods in Dehn surgery*, Proceedings of Knots 96, S. Suziki, ed., World Scientific Publishing (1997), 263–290.

[45] C. Gordon, *Boundary slopes of punctured tori in* 3*-manifolds*, Trans. Amer. Math. Soc., to appear.

[46] C. Gordon and R. Litherland, *Incompressible planar surfaces in* 3*-manifolds*, Topology Appl. **18** (1984), 121–144.

[47] C. Gordon and J. Luecke, *Only integral surgery can yield reducible manifolds*, Math. Proc. Cambr. Phil. Soc. **102** (1987), 94–101.

[48] C. Gordon and J. Luecke, *Knots are determined by their complements*, J. Amer. Math. Soc. **2** (1989), 371–415.

[49] C. Gordon and J. Luecke, *Surgeries producing essential tori, I*, Comm. Anal. Geom. **3** (1995), 597–644.

[50] C. Gordon and J. Luecke, *Reducible manifolds and Dehn surgery*, Topology **35** (1996), 385–410.

[51] C. Gordon and J. Luecke, *Dehn surgeries creating essential tori, II*, Comm. Anal. Geom., to appear.

[52] C. Gordon and J. Luecke, *Toroidal and boundary-reducing Dehn filling*, Topology Appl., to appear.

[53] C. Gordon and J. Luecke, *Annular and boundary reducing Dehn surgery*, Preprint.

[54] A. Hatcher, *On the boundary curves of incompressible surfaces*, Pacific J. Math. **99** (1982), 373–377.

[55] A. Hatcher and U. Oertel, *Boundary slopes for Montesinos knots*, Topology **28** (1989), 453–480.

[56] A. Hatcher and W. Thurston, *Incompressible surfaces in* 2*-bridge knots complements*, Invent. Math. **79** (1985), 225–246.

[57] C. Hayashi and K. Shimokawa, *Symmetric knots satisfy the cabling conjecture*, Math. Proc. Cambr. Phil. Soc., to appear.

[58] W. Heil, *Elementary surgery on Seifert fiber spaces*, Yokohama Math. J. **22** (1974), 135–139.

[59] J. Hempel, *Coverings of Dehn fillings of surface bundles*, Topology Appl. **24** (1986), 157–170.

[60] W. Jaco, *Lectures on Three-Manifold Topology*, CBMS Regional Conf. Ser. Math., Vol. 43 (1980).

[61] W. Jaco, *Adding a* 2*-handle to a* 3*-manifold: An application to property R*, Proc. Amer. Math. Soc. **92** (1984), 288–292.

[62] T. Kanenobu and H. Murakami, *Two-bridge knots with unknotting number* 1, Proc. Amer. Math. Soc. **98** (1986), 449–502.

[63] R. Kirby, *A calculus for framed links*, Invent. Math. **45** (1978), 35–56.

[64] R. Kirby, *Problems in low-dimensional topology*, Geometric Topology, Vol. 2, W. Kazez, ed., AMS/IP Studies in Advanced Mathematics (1996).

[65] R. Kirby and P. Melvin, *The* 3*-manifold invariants of Witten and Reshetikhin–Turaev for* $sl(2, \mathbf{C})$, Invent. Math. **105** (1991), 473–545.

[66] M. Lackenby, *Word hyperbolic Dehn surgery*, Preprint.

[67] C. Lescop, *Global Surgery Formula for the Casson–Walker Invariant*, Ann. Math. Studies, Vol. 140, Princeton University Press (1996).

[68] W.B.R. Lickorish, *A representation of orientable combinatorial* 3*-manifolds*, Ann. of Math. **76** (1962), 531–540.

[69] W.B.R. Lickorish, *The unknotting number of a classical knot*, Contemp. Math. **44** (1985), 117–121.

[70] R. Litherland, *Surgery on knots in solid tori, II*, J. London Math. Soc. (2) **22** (1980), 559–569.

[71] J. Luecke, *Dehn surgery on knots in the* 3*-sphere*, Proceedings of the International Congress of Mathematicians, Zürich, 1994, Birkhäuser, Switzerland (1995), 585–594.

[72] Y. Mathieu, *Sur des noeuds qui ne sont pas déterminés par leur compléments et problèmes de chirurgie dans les variétés de dimension* 3, Ph.D. thesis, Université de Marseille (1990).

[73] W. Menasco, *Closed incompressible surfaces in alternating knot and link complements*, Topology **23** (1984), 37–44.

[74] W. Menasco and M. Thistlethwaite, *Surfaces with boundary in alternating knot exteriors*, J. Reine Angew. Math. **426** (1992), 47–65.

[75] Y. Moriah and J.H. Rubinstein, *Heegaard structures of negatively curved manifolds*, Comm. Anal. Geom. **5** (1997), 375–412.

[76] L. Moser, *Elementary surgery along a torus knot*, Pacific J. Math. **38** (1971), 737–745.
[77] R. Naimi, *Constructing essential laminations in some 3-manifolds*, Pacific J. Math. **180** (1997), 153–186.
[78] W. Neumann and A. Reid, *Arithmetic of hyperbolic manifolds*, Topology '90, de Gruyter, Berlin (1992), 273–310.
[79] W. Neumann and D. Zagier, *Volumes of hyperbolic 3-manifolds*, Topology **24** (1985), 307–332.
[80] U. Oertel, *Closed incompressible surfaces in complements of star links*, Pacific J. Math. **111** (1984), 209–230.
[81] S. Oh, *Reducible and toroidal manifolds obtained by Dehn filling*, Topology Appl., to appear.
[82] C.D. Papakyriakopoulos, *On the asphericity of knots and Dehn's lemma*, Ann. of Math. **66** (1957), 1–26.
[83] Y. Rieck, *Heegaard structures of manifolds in the Dehn surgery space*, Preprint.
[84] R. Roberts, *Constructing taut foliations*, Comment. Math. Helv. **70** (1995), 516–545.
[85] D. Rolfsen, *Knots and Links*, 2nd edn, Publish or Perish (1990).
[86] D. Rolfsen, *Rational surgery calculus: Extension of Kirby's theorem*, Pacific J. Math. **110** (1984), 377–386.
[87] H. Rosenberg, *Foliations by planes*, Topology **7** (1968), 131–138.
[88] M. Scharlemann, *Unknotting number one knots are prime*, Invent. Math. **82** (1985), 37–55.
[89] M. Scharlemann, *Sutured manifolds and generalized Thurston norms*, J. Differential Geom. **29** (1989), 557–614.
[90] M. Scharlemann, *Producing reducible 3-manifolds by surgery on a knot*, Topology **29** (1990), 481–500.
[91] P. Shalen, *Representations of 3-manifold groups and applications in topology*, Proceedings of the International Congress of Mathematicians, Berkeley (1986), 607–614.
[92] P.B. Shalen, *Representations of 3-manifold groups*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 955–1044.
[93] D. Tanguay, *Chirurgie finie et noeuds rationnels*, Ph.D. thesis, UQAM (1995).
[94] R. Thom, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. **28** (1954), 17–86.
[95] W. Thurston, *The Geometry and Topology of 3-Manifolds*, Lecture Notes, Princeton University (1977).
[96] W. Thurston, *Three-dimensional manifolds, Kleinian groups, and hyperbolic geometry*, Bull. Amer. Math. Soc. **6** (1982), 357–381.
[97] W. Thurston, *Hyperbolic structure on 3-manifolds I: Deformation of acylindrical manifolds*, Ann. of Math. **124** (1986), 203–246.
[98] W. Thurston, *A norm for the homology of 3-manifolds*, Mem. Amer. Math. Soc. **339** (1986), 99–130.
[99] W. Thurston, *Three-Dimensional Geometry and Topology*, Vol. 1, Princeton University Press (1997).
[100] H. Tietze, *Über die topologischen Invarianten mehrdimensionaler Mannigfaltigkeiten*, Monatsh. Math. Phys. **19** (1908), 1–118.
[101] A. Tristram, *Some cobordism invariants for links*, Math. Proc. Camb. Phil. Soc. **66** (1969), 251–264.
[102] F. Waldhausen, *The word problem in fundamental groups of sufficiently large 3-manifolds*, Ann. of Math. **88** (1968), 272–280.
[103] F. Waldhausen, *Über Involutionen der 3-Sphare*, Topology **8** (1969), 81–91.
[104] A.H. Wallace, *Modifications and cobounding manifolds*, Can. J. Math. **12** (1960), 503–528.
[105] J. Weeks, Ph.D. thesis, Princeton University (1985).
[106] Y.-Q. Wu, *The reducibility of surgered manifolds*, Topology Appl. **43** (1992), 213–218.
[107] Y.-Q. Wu, *Essential laminations in surgered 3-manifolds*, Proc. Amer. Math. Soc. **115** (1992), 245–249.
[108] Y.-Q. Wu, *Incompressibility of surfaces in surgered 3-manifolds*, Topology **31** (1992), 271–279.
[109] Y.-Q. Wu, *Dehn surgery on arborescent knots*, J. Differential Geom. **42** (1996), 171–197.
[110] Y.-Q. Wu, *Dehn fillings producing reducible manifolds and toroidal manifolds*, Topology **37** (1998), 95–108.
[111] Y.-Q. Wu, *Sutured manifold hierarchies, essential laminations, and Dehn surgery*, J. Differential Geom., to appear.
[112] X. Zhang, *Unknotting number one knots are prime: a new proof*, Proc. Amer. Math. Soc **113** (1991), 611–612.

CHAPTER 5

Piecewise Linear Topology

John L. Bryant*

Florida State University, Tallahassee, FL 32306, USA

Contents

*Partially supported by NSF Grant DMS9300935.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

The piecewise linear category offers a rich structural setting in which to study many of the problems that arise in geometric topology. The first systematic accounts of the subject may be found in [2] and [63]. Whitehead's important paper [63] contains the foundation of the geometric and algebraic theory of simplicial complexes that we use today. More recent sources, such as [30], [50], and [66], together with [17] and [37], provide a fairly complete development of PL theory up through the early 1970s. This chapter will present an overview of the subject, drawing heavily upon these sources as well as others with the goal of unifying various topics found there as well as in other parts of the literature. We shall try to give enough in the way of proofs to provide the reader with a flavor of some of the techniques of the subject, while deferring the more intricate details to the literature. Our discussion will generally avoid problems associated with embedding and isotopy in codimension 2. The reader is referred to [12] for a survey of results in this very important area.

2. Basic definitions and terminology

Simplexes. A *simplex of dimension* p (a p*-simplex*) σ is the convex closure of a set of $(p+1)$ geometrically independent points $\{v_0, \ldots, v_p\}$ in Euclidean n-space $\mathbb{R}^n$. That is, each point x of σ can be expressed uniquely as $\sum t_i v_i$, where $0 \leqslant t_i \leqslant 1$ for $0 \leqslant i \leqslant p$ and $\sum t_i = 1$. (This is equivalent to requiring linear independence of the set of vectors $\{v_1 - v_0, \ldots, v_p - v_0\}$.) The v_i's are the *vertices* of σ; the t_i's are the *barycentric coordinates* of x in σ. We say that σ is *spanned* by its vertices, and write $\sigma = v_0 v_1 \cdots v_p$. The point $\beta(\sigma) = \sum \frac{1}{p+1} v_i$ is the *barycenter* of σ. A simplex τ spanned by a subset of the vertices is called a *face* of σ, written $\tau < \sigma$.

Simplicial complexes. A collection K of simplexes in $\mathbb{R}^n$ is called a (*simplicial*) *complex* provided

(1) if $\sigma \in K$ and $\tau < \sigma$, then $\tau \in K$,
(2) if $\sigma, \tau \in K$, then $\sigma \cap \tau < \sigma$ and $\sigma \cap \tau < \tau$, and
(3) K is *locally finite*; that is, given $x \in \sigma \in K$, then some neighborhood of x meets only finitely many τ in K. [1]

Simplicial complexes K and L are *isomorphic*, $K \cong L$, if there is a face-preserving bijection $K \leftrightarrow L$. The subset $|K| = \bigcup\{\sigma\colon \sigma \in K\}$ of $\mathbb{R}^n$ is called the *polyhedron* of K. Property (3) ensures that a subset of A of $|K|$ is closed in $|K|$ iff $A \cap \sigma$ is closed in σ for all $\sigma \in K$. That is, the weak topology on $|K|$ with respect to the collection K of simplexes coincides with the subspace topology on $|K|$. A complex L is a *subcomplex* of a complex K, $L < K$, provided $L \subseteq K$ and L satisfies (1)–(3). If $L < K$, then $|L|$ is a closed subset of $|K|$. If $A \subseteq |K|$ and $A = |L|$, for some $L < K$, we shall occasionally write $L = K|A$. For any complex K and any $p \geqslant 0$, we have the subcomplex $K^{(p)} = \{\sigma \in K\colon \dim \sigma \leqslant p\}$

[1] This is not a standard requirement, but we shall find it convenient for the purposes of this exposition. The astute reader, however, may notice an occasional lapse in our adherence to this restriction.

called the *p-skeleton of K*. For a simplex σ, the *boundary subcomplex* of σ is the subcomplex $\dot{\sigma} = \{\tau < \sigma\colon \tau \neq \sigma\}$. The *interior* of σ, $\mathring{\sigma} = \sigma - |\dot{\sigma}|$.

Subdivisions. A complex K_1 is a *subdivision of* K, $K_1 \prec K$, provided $|K_1| = |K|$ and each simplex τ of K_1 lies in some simplex σ of K. We write $(K_1, L_1) \prec (K, L)$ to denote that K_1 is a subdivision of K inducing a subdivision L_1 of L. If σ is a simplex, $L \prec \dot{\sigma}$, and $x \in \mathring{\sigma}$, then the subdivision $K = L \cup \{x w_0 w_1 \cdots w_k\colon w_0 w_1 \cdots w_k \in L\}$ is obtained from L by *starring* σ at x over L. A *derived* subdivision of K is one that is obtained by the following inductive process: assuming $K^{(p-1)}$ has been subdivided as a complex L and σ is a p-simplex of K, choose a point $\hat{\sigma}$ in $\mathring{\sigma}$ and star σ at $\hat{\sigma}$ over $L||\dot{\sigma}|$, thereby obtaining a subdivision of $K^{(p)}$. If we choose each $\hat{\sigma} = \beta(\sigma)$, the resulting derived subdivision is called the *first barycentric subdivision of* K and is denoted by K^1. More generally, K^r will note the rth-barycentric subdivision of K: $K^r = (K^{r-1})^1$ (where $K^0 = K$). There are relative versions of this process: if L is a subcomplex of K, inductively choose points $\hat{\sigma} \in \operatorname{int}\sigma$ for $\sigma \notin L$. The result is a *derived subdivision of* K mod L.

A subcomplex L of a complex K is *full* in K, $L \lhd K$, if a simplex σ of K belongs to L whenever all of its vertices are in L. If L is a subcomplex of K and K' is a derived subdivision of K, then the subcomplex L' of K' subdividing L is full in K'.

Any two subdivisions $L \prec K$ and $J \prec K$ have a common subdivision. The set $\mathcal{C} = \{\sigma \cap \tau\colon \sigma \in L,\ \tau \in J\}$ is a collection of convex linear cells that forms a *cell complex*: given $C, D \in \mathcal{C}$, $C \cap D \in \mathcal{C}$ is a face of each. $\mathcal{C}$ can be subdivided into simplexes by induction using the process described above. $\mathcal{C}$ can also be subdivided into simplexes without introducing any additional vertices, other than those in the convex cells of $\mathcal{C}$, by a similar process: order the vertices of $\mathcal{C}$ and, assuming the boundary of a convex cell C of $\mathcal{C}$ has been subdivided, choose the first vertex v of C and form simplexes $v w_0 \cdots w_p$ where $w_0 \cdots w_p$ is a simplex in the boundary of C not containing v. A consequence of the latter construction is that if subdivisions $L \prec K$ and $J \prec K$ share a common subcomplex M, then a common subdivision of L and J can be found containing M as a subcomplex. Finally, if $L < K$ and $L' \prec L$, then there is a subdivision $K' \prec K$ such that $K'||L| = L'$: proceed inductively starring a p-simplex σ of K not in L at an interior point x over $K^{(p-1)'}||\dot{\sigma}|$.

Stars and links. Given a complex K and a simplex $\sigma \in K$, the *star* and *link* of σ in K are the subcomplexes $\mathrm{St}(\sigma, K) = \{\tau \in K\colon \text{for some } \eta \in K,\ \sigma, \tau < \eta\}$, and $\mathrm{Lk}(\sigma, K) = \{\tau \in \mathrm{St}(\sigma, K)\colon \tau \cap \sigma = \emptyset\}$, respectively. We let $\mathrm{st}(\sigma, K) = |\mathrm{St}(\sigma, K)|$ and $\mathrm{lk}(\sigma, K) = |\mathrm{Lk}(\sigma, K)|$. The *open star* of σ in K, $\mathring{\mathrm{st}}(\sigma, K) = \mathrm{st}(\sigma, K) - \mathrm{lk}(\sigma, K)$. One can easily show that the collection $\{\mathring{\mathrm{st}}(v, K^r)\colon v \in (K^r)^{(0)},\ r = 0, 1, \ldots\}$ forms a basis for the open sets in $|K|$.

Simplicial and piecewise linear maps. Given complexes K and L, a *simplicial map* $f\colon K \to L$ is a map (we still call) $f\colon |K| \to |L|$ such that for each $\sigma \in K$, $f|\sigma$ maps σ linearly onto a simplex of L. A simplicial map $f\colon K \to L$ is *nondegenerate* if $f|\sigma$ is injective for each $\sigma \in K$. A simplicial map is then determined by its restriction to the vertices of K. A map $f\colon |K| \to |L|$ is *piecewise linear* or PL if there are subdivisions $K' \prec K$ and $L' \prec L$ such that $f\colon K' \to L'$ is simplicial. Polyhedra $|K|$ and $|L|$ are *piecewise linearly* (or PL) *homeomorphic*, $|K| \cong |L|$, if they have subdivisions $K' \prec K$ and $L' \prec L$ such that $K' \cong L'$.

SIMPLICIAL APPROXIMATION THEOREM. *If K and L are complexes and $f : |K| \to |L|$ is a continuous function, then there is a subdivision $K' \prec K$ and a simplicial map $g : K' \to L$ homotopic to f. Moreover, if $\varepsilon : |L| \to (0, \infty)$ is continuous, then there are subdivisions $K' \prec K$ and $L' \prec L$ and a simplicial map $g : K' \to L'$ such that g is ε-homotopic to f; that is, there is a homotopy $H : |K| \times [0, 1] \to |L|$ from f to g such that* $\operatorname{diam}(H(x \times [0, 1])) < \varepsilon(f(x))$ *for all* $x \in |K|$.

The proof of this theorem is elementary and can be found in a number of texts. (See, for example, [45] or [54].) The idea of the proof is to get an r such that for each vertex v of K^r, $f(\operatorname{st}(v, K^r)) \subseteq \mathring{\operatorname{st}}(w, L)$ for some vertex w of L. The assignment of vertices $v \mapsto w$, defines a vertex map $g : (K^r)^{(0)} \to L^{(0)}$ that extends to a simplicial map $g : K^r \to L$ homotopic to f (by a straight line homotopy). This can be done whenever K is finite. When K is not finite, one may use a *generalized barycentric subdivision* of K, constructed inductively as follows. Assuming J is a subdivision of $K^{(p-1)}$, and σ is a p-simplex of K, let K_σ be the subdivision of σ obtained by starring σ at $\beta(\sigma)$ over $J||\dot{\sigma}|$. Let $n = n_\sigma$ be a non-negative integer, and let K'_σ be the nth-barycentric subdivision of K_σ mod $L||\dot{\sigma}|$. It can be shown that any open cover of $|K|$ can be refined by $\{\operatorname{st}(v, K') \colon v \in K'^{(0)}\}$ for some generalized barycentric subdivision K' of K.

To get the "moreover" part, start with a (generalized) rth barycentric subdivision L' of L such that vertex stars have diameter less than ε.

Generalized barycentric subdivisions can also be used to show that if U is an open subset of the polyhedron $|K|$ of a complex K, then U is the polyhedron of a complex J each simplex of which is linearly embedded in a simplex of K.

Combinatorial manifolds. A *combinatorial n-manifold* is a complex K for which the link of each p-simplex is PL homeomorphic to either the boundary of an $(n - p)$-simplex or to an $(n - p - 1)$-simplex. If there are simplexes of the latter type, they constitute a subcomplex ∂K of K, the *boundary* of K, which is, in turn, a combinatorial $(n - 1)$-manifold without boundary. If K is a combinatorial n-manifold, then $|K|$ is a topological n-manifold, possibly with boundary $|\partial K|$.

Triangulations. A *triangulation* of a topological space X consists of a complex K and a homeomorphism $t : |K| \to X$. Two triangulations $t : |K| \to X$ and $t' : |K'| \to X$ of X are *equivalent* if there is a PL homeomorphism $h : |K| \to |K'|$ such that $t' \circ h = t$. A PL n-manifold is a space (topological n-manifold) M, together with a triangulation $t : |K| \to M$, where K is a combinatorial n-manifold. Such a triangulation will be called a PL triangulation of M or a *PL structure* on M. $\partial M = |\partial K|$ and $\operatorname{int} M = M - \partial M$. M is PL *n-ball* (respectively, PL *n-sphere*) if we can choose K to be an n-simplex (respectively, the boundary subcomplex of an $(n + 1)$-simplex). In a similar manner we may define a triangulation $K > L$ of a pair $X \supset Y$, where Y is closed in X (or for a triad $X \supset Y, Z$, or n-ad, etc.).

A PL structure on a topological n-manifold M can also be prescribed by an *atlas* Σ on M, consisting of a covering $\mathcal{U}$ of open sets (*charts*) in M together with embeddings $\phi_U : U \to \mathbb{R}^n$, $U \in \mathcal{U}$, such that if $U, V \in \mathcal{U}$, then $\phi_V(\phi_U)^{-1} : \phi_U(U \cap V) \to \phi_V(U \cap V)$ is piecewise linear. Here we assume that open subsets of $\mathbb{R}^n$ inherit triangulations from linear triangulations of $\mathbb{R}^n$ as described above. Two atlases Σ and Σ' are *equivalent* if there is a (topological) homeomorphism $h : M \to M$ such that the union of Σ and $h(\Sigma')$ forms an atlas, where $h(\Sigma')$ is the atlas consisting of the cover $\{h(U')\colon U' \in \mathcal{U}'\}$ and embeddings $\phi_{U'}h^{-1}$. An atlas Σ on M determines a PL triangulation of M as follows. If M is compact, cover M by a finite number of compact polyhedra obtained from a finite cover of open sets in Σ, and triangulate inductively. If M is not compact, then dimension theory provides a cover $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_n$, subordinate to $\mathcal{U}$, such that the members of $\mathcal{X}_i$, $i = 0, 1, \ldots, n$, are mutually exclusive, compact polyhedra. One can then proceed as in the compact case. It is not difficult to show that atlases Σ and Σ' are equivalent if, and only if, the induced triangulations of M are equivalent.

One may also consider the problem of "triangulating" a diagram of polyhedra and PL maps; that is, subdividing all spaces so that each of the mappings in the diagram is simplicial. If the diagram forms a "one-way tree" in which each polyhedron is compact and is the domain of at most one mapping, then it is possible to use an inductive argument, based on the following construction, to triangulate the diagram. Given a simplicial mapping $f : K \to L$ and a subdivision $L' \prec L$, form the cell complex $\mathcal{C} = \{\sigma \cap f^{-1}(\tau)\colon \sigma \in K, \tau \in L'\}$, and subdivide $\mathcal{C}$ as a simplicial complex K' without introducing any new vertices, as above. Then $f : K' \to L'$ is simplicial.

If a diagram does not form a one-way tree, then it may not be triangulable, as a simple example found in [66] illustrates. Let $|K| = [-1, 1]$, $|L| = |J| = [0, 1]$, let $f : |K| \to |L|$ be defined by $f(x) = |x|$, and let $g : |K| \to |J|$ be defined by $g(x) = x$, if $x \geqslant 0$, and $g(x) = -x/2$, if $x \leqslant 0$. The problem is that there is a sequence $\{1/2, 1/4, 1/8, \ldots\}$ in $|L|$ such that $gf^{-1}(1/2^i) \cap gf^{-1}(1/2^{i+1}) \neq \emptyset$. In [9] it is shown that a two-way diagram

$$|J| \xleftarrow{g} |K| \xrightarrow{f} |L|$$

can be triangulated provided it does not admit such sequences. (See [9] for a precise statement of the theorem and its proof.)

The PL category. The *piecewise linear category*, PL, can now be described as the category whose objects are triangulated spaces, or, simply, polyhedra, and whose morphisms are PL maps. The usual Cartesian product and quotient constructions can be carried out in PL with some care: the Cartesian product of two polyhedra doesn't have a well-defined triangulation (since the product of two simplexes is rarely a simplex), and a complex obtained by an identification on the vertices of another complex may not give a complex with the expected (or desired) polyhedron. For example, identifying the vertices of a 1-simplex will not produce a complex with polyhedron homeomorphic to S^1, since the only simplicial map from a 1-simplex making this identification is a constant map. One must first subdivide the simplex (it takes two derived subdivisions). Either of the two processes described above for finding a common subdivision of two subdivisions of a complex may be used to triangulate the Cartesian product of two complexes K and L. For example, one can inductively star the convex cells $\sigma \times \tau$ ($\sigma \in K, \tau \in L$). Alternatively, one can order $K^{(0)} \times L^{(0)}$, perhaps using a lexicographic ordering resulting from an ordering of $K^{(0)}$ and $L^{(0)}$ separately, and inductively triangulate the convex linear cells $\sigma \times \tau$ ($\sigma \in K,\ \tau \in L$) without introducing any new vertices.

Joins: Cones and suspensions. The join operation is a more natural operation in PL than are products and quotients. Disjoint subsets A and B in $\mathbb{R}^n$ are *joinable* provided any two line segments joining points of A to points of B meet in at most a common endpoint (or coincide). If A and B are joinable then the *join* of A and B, $A * B$, is the union of all line segments joining a point of A to a point of B. We can always "make" A and B joinable: if $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$, then $A \times 0 \times 0$ and $0 \times B \times 1$ are joinable in $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{m+n+1}$. We assume the convention that $A * \emptyset = \emptyset * A = A$. If $A \cap B = C \neq \emptyset$, then A and B are *joinable relative to* C if $A - C$ and $B - C$ are joinable and every line segment joining a point of $A - C$ and $B - C$ misses C. Then $A * B(\operatorname{rel} C) = [(A - C) * (B - C)] \cup C$ denotes the *reduced join of A and B relative to C*. For example, given a simplex $\sigma = v_0 \cdots v_p$ and faces $\tau = v_0 \cdots v_i$ and $\eta = v_j \cdots v_p$ with $j \leqslant i + 1$, then $\sigma = \tau * \eta(\operatorname{rel} \tau \cap \eta)$. Likewise, if K and L are finite complexes in $\mathbb{R}^n$ such that $|K|$ and $|L|$ are joinable, then we can define the *join complex*, $K * L = \{\sigma * \tau \subseteq \mathbb{R}^{m+n}\colon\ \sigma \in K$ and $\tau \in L\}$. For example, if σ is a simplex in a complex K, then $\operatorname{St}(\sigma, K) = \sigma * \operatorname{Lk}(\sigma, K)$. Unlike the case for products and quotients, triangulations of compact spaces X and Y induce a canonical triangulation of $X * Y$. An important artifact of the join construction is that the join of two spaces $A * B$ comes equipped with a *join parameter* obtained from a natural map $s : A * B \to [0, 1]$ that maps each line segment in $A * B$ from a point of A to a point of B linearly onto $[0, 1]$. When K and L are finite complexes, the map s is a simplicial map from $K * L$ onto the simplex $[0, 1]$. With the aid of the join parameter, one can easily extend simplicial maps $f : H \to K$ and $g : J \to L$ between finite complexes to their joins, $f * g : H * J \to K * L$.

Two special cases of the join construction are the cone and suspension. Given a compact set X and a point v, the *cone* on X with vertex v, $C(X, v) = v * X$. We may also write $C(X)$ to denote $C(X, v)$. The *suspension* of X, $\Sigma(K) = S^0 * K$, where S^0 is the 0-sphere. One defines cone and suspension complexes of a (finite) complex K similarly. As observed above, if v is a vertex of a complex K, then $\operatorname{St}(v, K) \cong v * \operatorname{Lk}(v, K)$. Using the join construction for simplicial maps, one can easily prove PL equivalence for stars of vertices.

THEOREM 2.1. *Suppose that X is a polyhedron, $x \in X$, and K_1 and K_2 are equivalent triangulations of X containing x as a vertex. Then* $\operatorname{st}(x, K_1) \cong \operatorname{st}(x, K_2)$.

PROOF. Without loss of generality we may assume that K_2 is a subdivision of K_1 so that $\operatorname{st}(x, K_2) \subseteq \operatorname{st}(x, K_1)$. Hence, for each point y of $\operatorname{lk}(x, K_2)$, there is a unique point $z \in \operatorname{lk}(x, K_1)$ such that $y \in x * z \subseteq x * \operatorname{lk}(x, K_1) = \operatorname{st}(x, K_1)$. Conversely, for each $z \in \operatorname{lk}(x, K_1)$ there is a unique $y \in \operatorname{lk}(x, K_2)$ such that $x * z \cap \operatorname{lk}(x, K_2) = y$. Moreover, if z is a vertex of $\operatorname{Lk}(x, K_1)$, then y is a vertex of $\operatorname{Lk}(x, K_2)$. Thus, we can get a simplicial isomorphism f from $\operatorname{Lk}(x, K_2)$ to a subdivision $\operatorname{Lk}(x, K_1)'$ of $\operatorname{Lk}(x, K_1)$ by extending the map above from the vertices of $\operatorname{Lk}(x, K_2)$ into $\operatorname{lk}(x, K_1)$. Extending further to the cones over x gives the desired equivalence. □

As pointed out in [66] and [50], the natural projection $\mathrm{lk}(x, K_2) \to \mathrm{lk}(x, K_1)$ along cone lines is not linear on the simplexes of K_2, although it does match up the simplexes of $\mathrm{Lk}(x, K_2)$ with those of the subdivision $\mathrm{Lk}(x, K_1)'$ of $\mathrm{Lk}(x, K_1)$. (This is the "Standard Mistake".)

The proof of the following important theorem can be found in [50].

THEOREM 2.2. *Suppose B^p (respectively, S^p) denotes a PL ball (respectively, sphere) of dimension p, then*
(i) $B^p * B^q = B^{p+q+1}$,
(ii) $B^p * S^q = B^{p+q+1}$, *and*
(iii) $S^p * S^q = S^{p+q+1}$.

For example, if K is a combinatorial n-manifold and σ is a p-simplex of K, then $\mathrm{st}(\sigma, K) \cong \sigma * \mathrm{lk}(\sigma, K) \cong B^n$.

An elementary argument shows that the join operation is associative. This implies, for example, that a k-fold suspension $\Sigma^k(X) = \Sigma(\Sigma(\cdots(\Sigma(X))\cdots))$ of a compact polyhedron X is PL homeomorphic to $S^{k-1} * X$. The proof of the following proposition is a pleasant exercise in the use of some of the ideas presented so far.

PROPOSITION 2.3. *If X is a compact polyhedron, then*

$$C(X) \times [-1,1] \cong C\big((X \times [-1,1]) \cup \big(C(X) \times \{-1,1\}\big)\big)$$

by a homeomorphism that preserves $C(X) \times [-1,0]$ and $C(X) \times [0,1]$. In particular, if $J > J_+, J_-, J_0$ is a triangulation of $C(X) \times [-1,1] \supset C(X) \times [0,1]$, $C(X) \times [-1,0]$, $C(X) \times \{0\}$, then

$$\begin{aligned}&\big(\mathrm{st}(v,J), \mathrm{st}(v,J_+), \mathrm{st}(v,J_-), \mathrm{st}(v,J_0)\big)\\ &\quad\cong \big(C(X) \times [-1,1], C(X) \times [0,1], C(X) \times [-1,0], C(X) \times \{0\}\big)\end{aligned}$$

(where v is the vertex of $C(X)$).

Proposition 2.3 in turn may be applied to give a proof of a PL version of Morton Brown's Collaring Theorem [7] . A subpolyhedron Y of a polyhedron X is *collared* in X if Y has a neighborhood in X PL homeomorphic to $Y \times I$. Y is *locally collared* in X if each $x \in Y$ has a neighborhood pair (U, V) in (X, Y) such that $(U, V) \cong (V \times I, V \times \{0\})$.

THEOREM 2.4. *If the subpolyhedron Y of X is locally collared in X, then Y is collared in X.*

PROOF. Let $K > L$ be a triangulation of $X \supset Y$, and assume that for each vertex $v \in L$, $\mathrm{st}(v, L)$ lies in a collared neighborhood. That is, v has a neighborhood pair (U, V) PL homeomorphic to $(\mathrm{st}(v, L) \times I, \mathrm{st}(v, L) \times \{0\})$ $(I = [0, 1])$. By Proposition 2.3, we may assume that $U = \mathrm{st}(v, K)$. Let $X^+ = X \cup_{Y \times \{0\}} (Y \times [-1, 0])$. Then $U \cup_{V \times \{0\}} (V \times [-1, 0]) \cong V \times [-1, 1]$ is a neighborhood of $v = (v, 0)$ in X^+, and $V \times [-1, 1] \cong v *$

$(\text{lk}(v, L) \times [-1, 1] \cup V \times \{-1, 1\})$. Let $\Sigma = \text{lk}(v, L) \times [-1, 1] \cup V \times \{-1, 1\}$, and let $v' = (v, -\frac{1}{2})$. Then there is a homeomorphism $h_v : V \times [-1, 1] \to V \times [-1, 1]$ such that $h_v(v) = v'$, $h_v|\Sigma = \text{id}$, and h_v sends each $v * z$, $z \in \Sigma$, "linearly" onto $v' * z$. In particular, h_v commutes with the projection map $V \times [-1, 1] \to V$.

Now let $K' > L'$ be a derived subdivision of $K > L$. Write $L'^{(0)} = V_0 \cup V_1 \cup \cdots \cup V_m$, where $V_i = \{\hat{\sigma} \in L' \colon \dim \sigma = i\}$. Then for $v_1, v_2 \in V_i$ $\text{st}(v_1, K') \cap \text{st}(v_2, K') \subseteq \text{lk}(v_1, K') \cap \text{lk}(v_2, K')$ so that h_{v_1} is the identity on $\text{st}(v_1, K') \cap \text{st}(v_2, K')$. Let $h_i : X^+ \to X^+$ be the PL homeomorphism satisfying $h_i = h_v$ on $\text{st}(v, K') \cup_{\text{st}(v, L') \times \{0\}} (\text{st}(v, L') \times [-1, 0])$ for $v \in V_i$ and $h_i = \text{id}$, otherwise. Then $h = h_m \circ \cdots \circ h_1 \circ h_0 : X^+ \to X^+$ is a homeomorphism that takes $(Y \times [-1, 0], Y \times \{0\})$ to $(Y \times [-1, -\frac{1}{2}], Y \times \{-\frac{1}{2}\})$. Hence, h^{-1} takes $Y \times [-\frac{1}{2}, 0]$ onto a neighborhood of Y in X. □

COROLLARY 2.5. *Suppose that X is a PL n-manifold with boundary Y. Then Y is collared in X.*

PROOF. Each $x \in Y$ has a neighborhood N such that $N \cong B^n$ and $N \cap Y \cong B^{n-1}$. Since S^{n-1} is collared in B^n, x has a neighborhood PL homeomorphic to $B^{n-1} \times [0, 1]$. □

Join structures play an essential role in PL theory. They lie at the heart of many constructions and much of the structure theory. We conclude this section with three important examples.

Simplicial mapping cylinders. Suppose $f : K \to L$ is a simplicial map. (If K is not finite, assume additionally that $f^{-1}(v)$ is a finite complex for each vertex v of L.) Choose first derived subdivisions K' of K and L' of L such that $f : K' \to L'$ is still simplicial. For example, we can choose $L' = L^1$, the first barycentric subdivision of L, and for each $\sigma \in K$, choose a point $\hat{\sigma} \in \mathring{\sigma} \cap f^{-1}(\beta(f(\sigma)))$ at which to star σ. The *simplicial mapping cylinder of f* is the subcomplex of $L' * K'$,

$$C_f = \{\hat{\tau}_1 \hat{\tau}_2 \cdots \hat{\tau}_j * \hat{\sigma}_1 \hat{\sigma}_2 \cdots \hat{\sigma}_i \mid \tau_1 < \cdots < \tau_j < f(\sigma_1), \sigma_1 < \cdots < \sigma_i \in K\} \cup L'.$$

Thus, a simplex of C_f is either in L' or is of the form $\alpha * \beta \in L' * K'$, where, for some $\tau \in L$ and $\sigma \in K$, $\alpha \subseteq \tau$, $\beta \subseteq \sigma$, and $\tau < f(\sigma)$. There is a natural projection $\gamma : C_f \to L$ defined by

$$\gamma(\hat{\tau}_1 \hat{\tau}_2 \cdots \hat{\tau}_j * \hat{\sigma}_1 \hat{\sigma}_2 \cdots \hat{\sigma}_i) = \hat{\tau}_1 \hat{\tau}_2 \cdots \hat{\tau}_j f(\hat{\sigma}_1) f(\hat{\sigma}_2) \cdots f(\hat{\sigma}_i).$$

Figure 1 illustrates the simplicial mapping cylinder of a simplicial map $f : \sigma \to \tau$ from a 2-simplex σ to a 1-simplex τ.

As is shown in [15], the simplicial mapping cylinder C_f is topologically homeomorphic to the topological mapping cylinder $|K| \times I \cup_{f \times \{1\}} |L|$. If f is degenerate, however, any PL map $(|K| \times I) \coprod |L| \to C_f$ restricting to f on $|K| \times 1$ will fail to be one-to-one on $|K| \times [0, 1)$.

If $f : K \to L$ is the identity on a complex $H < K \cap L$, then one can also define the *reduced* simplicial mapping cylinder, a subcomplex of $L' * K'$ rel H, where L' and K' are

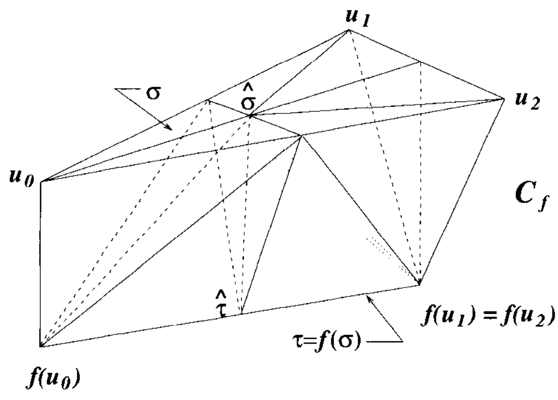


Fig. 1.

first derives mod H:

$$C_f \operatorname{rel} H = \{\alpha * \hat{\tau}_1\hat{\tau}_2 \cdots \hat{\tau}_j * \hat{\sigma}_1\hat{\sigma}_2 \cdots \hat{\sigma}_i \mid \alpha < \tau_1 < \tau_2 < \cdots < \tau_j < f(\sigma_1),\ \alpha \in H,\ \tau_1 \in L - H, \sigma_1 < \sigma_2 < \cdots < \sigma_i\} \cup L'.$$

Suppose $f: X \to Y$ is a PL mapping between polyhedra. In light of the comment above, we may refer to "the" PL mapping cylinder M_f of f, obtained from triangulations K of X and L of Y under which $f: K \to L$ is simplicial. M_f is well-defined topologically, but its combinatorial structure will depend on K and L. If $f|A$ is an embedding for some subpolyhedron A of X, we may also define the reduced PL mapping cylinder M_f rel A.

Dual subcomplexes. Given complexes $L < K$, let K' be the first barycentric subdivision of K mod L, and let $J = \{\sigma \in K' \mid \sigma \cap |L| = \emptyset\}$. Then J is the *dual* of L in K. In particular, if K is an n-complex and if $L = K^{(p)}$ is the p-skeleton of K, then J is called the *dual* $(n-p-1)$*-skeleton* of K, and is denoted by $\widetilde{K}^{(n-p-1)}$. Whenever J is the dual of L in K, K' is isomorphic to a subcomplex of $L * J$, since every simplex of K' is either in L, in J, or is the join of a simplex of L and a simplex of J. It is occasionally useful to consider relative versions of duals. For example, if K is a combinatorial n-manifold with boundary ∂K, then the dual $(n-p-1)$-skeleton of K rel ∂K is the dual of $K^{(p)} \cup \partial K$.

Dual cell structures. Suppose K is a combinatorial n-manifold (possibly with boundary), and K' is a first derived subdivision. Given a p-simplex σ in K, $K'|\operatorname{lk}(\sigma, K)$ is naturally isomorphic to the subcomplex $\widetilde{K}_\sigma = \{\hat{\tau}_1\hat{\tau}_2 \cdots \hat{\tau}_m: \sigma < \tau_1 < \cdots < \tau_m \in K,\ \sigma \neq \tau_1\}$ of K'. Thus, $|\widetilde{K}_\sigma| \cong S^{n-p-1}$ or B^{n-p-1}, and, hence, $\widetilde{B}_\sigma = \hat{\sigma} * |\widetilde{K}_\sigma|$ is a PL $(n-p)$-ball. $\widetilde{B}_\sigma$ is the *dual cell* to σ, and the collection $\widetilde{K}$ of dual cells is called the *dual cell complex* of K. $\widetilde{K}$ satisfies the conditions: $\sigma < \tau$, whenever $\widetilde{B}_\tau \subseteq \partial \widetilde{B}_\sigma$, and

$$\widetilde{B}_\sigma \cap \widetilde{B}_\tau = \begin{cases} \widetilde{B}_\eta, & \text{if } \eta = \sigma * \tau \ (\operatorname{rel} \sigma \cap \tau) \text{ is a simplex of } K, \text{ and} \\ \emptyset, & \text{otherwise.} \end{cases}$$

Figure 2 illustrates cell-dual cell pairs for a 1-dimensional face σ of a 2-simplex τ.

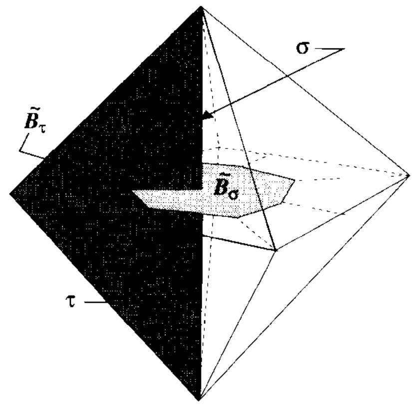


Fig. 2.

3. Regular neighborhoods

Derived neighborhoods. Given a subcomplex L of a complex K, the *simplicial neighborhood* of L in K is the subcomplex

$$\begin{aligned} N(L, K) &= \{\sigma\colon \sigma \in K,\ \sigma < \tau,\ \tau \cap |L| \neq \emptyset\} \\ &= \bigcup \{\mathrm{St}(v, K)\colon v \in L^{(0)}\}. \end{aligned}$$

Suppose $L \lhd K$. Let $C(L, K) = \{\sigma \in K\colon \sigma \cap |L| = \emptyset\}$, the *simplicial complement* of L in K, and let K' be a derived subdivision of K mod $L \cup C(L, K)$. Then $N(L, K')$ is a *derived neighborhood* of L in K. Any two derived neighborhoods corresponding to derived subdivisions K_1 and K_2 of K mod $L \cup C(L, K)$ are canonically isomorphic via an isomorphism $\phi : K_1 \to K_2$ that is the identity on $L \cup C(K, L)$. The *boundary* of $N(L, K')$ is the subcomplex $\dot{N}(L, K') = \{\sigma \in N(L, K')\colon \sigma \cap |L| = \emptyset\}$. Given $\varepsilon > 0$, the ε-neighborhood of L in K is a derived neighborhood constructed as follows. Since $L \lhd K$, the simplicial map $f : K \to [0, 1]$ defined by the vertex map

$$f(v) = \begin{cases} 0, & \text{if } v \in L, \\ 1, & \text{if } v \notin L, \end{cases}$$

has the property that $f^{-1}(0) = L$. For any simplex σ of K such that $\sigma \notin L \cup C(L, K)$ choose $\hat{\sigma} \in \mathring{\sigma} \cap f^{-1}(\varepsilon)$. Let K' be the resulting derived subdivision of K mod $L \cup C(L, K)$, and set $N_\varepsilon(L, K) = N(L, K')$.

EXAMPLE. Given a complex K and $p \geqslant 0$, let $L = K^{(p)}$, let K' be the first barycentric subdivision of K mod L, let $\widetilde{L} < K'$ be the dual of L, and let K'' be a derived subdivision of K' mod $L \cup \widetilde{L}$. Then $N(L, K'') \cup N(\widetilde{L}, K'') = K''$ and $\dot{N}(L, K'') = \dot{N}(\widetilde{L}, K'')$.

PROPOSITION 3.1. *Suppose $L \lhd K$ and $(K_1, L_1) \prec (K, L)$. Then there are derived neighborhoods $N(L, K')$ and $N(L_1, K_1')$ such that $|N(L, K')| = |N(L_1, K_1')|$.*

PROOF. Given $f : K \to [0, 1]$ as above, choose $\varepsilon > 0$ so that $f^{-1}((0, \varepsilon))$ contains no vertex of K or K_1. For each simplex σ of K (respectively, K_1) that meets $|L| (= |L_1|)$, choose $\hat{\sigma} \in \mathring{\sigma} \cap f^{-1}(\varepsilon)$. □

Regular neighborhoods. Given polyhedra $Y \subseteq X$, choose a triangulation K of X containing a subcomplex L triangulating Y. By passing to a derived subdivision of K mod L, we may assume that $L \lhd K$. The polyhedron $N = |N(L, K')|$ is called a *regular neighborhood* of Y in X. Proposition 3.1 can be applied to prove the following uniqueness theorem.

THEOREM 3.2. *Suppose N_1 and N_2 are regular neighborhoods of Y in X. Then there is a PL homeomorphism $h : X \to X$ such that $h|Y = \mathrm{id}$ and $h(N_1) = h(N_2)$. If Y is compact, then we can choose h so that h is the identity outside a compact subset of X.*

PROOF. Suppose $N_1 = |N(L_1, K_1')|$ and $N_2 = |N(L_2, K_2')|$, where $K_i > L_i$ triangulates $X \supset Y$, and $L_i \lhd K_i$. Let $K_0 > L_0$ be a triangulation of $X \supset Y$ subdividing both K_1 and K_2. By Proposition 3.1 there are derived subdivisions K_i'' of K_i mod $L_i \cup C(L_i, K_i)$, $i = 0, 1, 2$, such that $|N(L_0, K_0'')| = |N(L_1, K_1'')| = |N(L_2, K_2'')|$. By the canonical uniqueness, there are isomorphisms $\phi_i : K_i' \to K_i''$, fixed on $L_i \cup C(L_i, K_i)$, $i = 1, 2$, taking $N(L_i, K_i')$ to $N(L_i, K_i'')$. The composition $\phi_2^{-1} \circ \phi_1$ is a PL homeomorphism of X that is the identity on $Y \cup [C(L_1, K_1) \cap C(L_2, K_2)]$ and takes N_1 to N_2. □

THEOREM 3.3. *Suppose that N is a regular neighborhood of Y in X. Then a regular neighborhood of $\dot{N}$ in X is PL homeomorphic to $\dot{N} \times I$.*

PROOF. Suppose $N = |N(L, K')|$ where $K > L$ triangulates $X \supset Y$ and $L \lhd K$. Without loss of generality, $N = |N(L, K')| = |N_{1/2}(L, K')| = f^{-1}([0, 1/2])$, where $f : K \to [0, 1]$ is the simplicial map described above. For any simplex $\sigma \in K - (L \cup C(L, K))$, $f^{-1}([1/4, 3/4]) \cap \sigma$ is canonically PL homeomorphic to $f^{-1}(1/2) \times [1/4, 3/4]$. These homeomorphisms fit together naturally to give the desired result. □

The next theorem follows easily from Theorems 2.4 and 3.2.

THEOREM 3.4. *Suppose Y is a subpolyhedron of a polyhedron X such that Y is locally collared in X. Then a regular neighborhood of Y in X is a collar.*

THEOREM 3.5. *Suppose that N is a regular neighborhood of Y in X. Then N is PL homeomorphic to the mapping cylinder C_ϕ of a PL map $\phi : \dot{N} \to Y$.*

PROOF. As above, we suppose $K > L$ triangulates $X \supset Y$ with $L \lhd K$ and $N = |N(L, K')| = |N(L', K')| = |N_{1/2}(L', K')| = f^{-1}([0, 1/2])$, where $K' > L'$ is a first derived subdivision of $K > L$ mod $C(L, K)$. Any simplex $\sigma \in K - (L \cup C(L, K))$ is a join, $\sigma = \tau * \eta$, with $\tau \in L$ and $\eta \in C(L, K)$. The vertex assignment $\widehat{\tau * \eta} \mapsto \hat{\tau}$ defines a simplicial map $\phi : \dot{N}(L', K') \to L'$, and $C_\phi = N(L', K')$. □

The proof of the following theorem is left as an exercise. (Use Theorem 3.3.)

THEOREM 3.6. *Suppose X is a subpolyhedron of a PL manifold M. Then a regular neighborhood N of X in M is a PL manifold. If X is in the interior of M and $N = |N(L, K')|$ for some triangulation $K > L$ of $M \supset X$, then $\partial N = |\dot{N}(L, K')|$.*

A converse of Theorem 3.6 is contained in the following Simplicial Neighborhood Theorem. We state the theorem along with a selection of some of its more important corollaries. A proof may be found in [50].

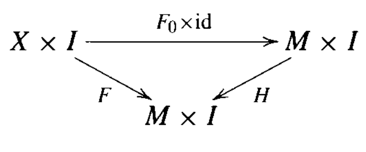
THEOREM 3.7 (Simplicial Neighborhood Theorem). *Suppose X is a subpolyhedron in the interior of a PL manifold M, and N is a neighborhood of X in M. Then N is a regular neighborhood of X if and only if*
(i) *N is a PL manifold with boundary, and*
(ii) *there is a triangulation $K > L, J$ of $N \supset X, \partial N$ with $L \lhd K$, $K = N(L, K)$ and $J = \dot{N}(L, K)$.*

COROLLARY 3.8. *If $B^n \subseteq S^n$ is a PL ball in a PL sphere, then $C\ell(S^n - B^n) \cong B^n$.*

COROLLARY 3.9. *If $N_1 \subseteq \operatorname{int} N_2$ are two regular neighborhoods of X in $\operatorname{int} M$, then $C\ell(N_2 - N_1) \cong \partial N_1 \times I$.*

COROLLARY 3.10 (Combinatorial Annulus Theorem). *If B_1 and B_2 are PL n-balls with $B_1 \subseteq \operatorname{int} B_2$, then $C\ell(B_2 - B_1) \cong S^{n-1} \times I$.*

The Regular Neighborhood Theorem. The Regular Neighborhood Theorem provides a strong isotopy uniqueness theorem for regular neighborhoods of X in M. Given a subpolyhedron X of a polyhedron M, an *isotopy* of X in M is a level-preserving, closed, PL embedding $F : X \times I \to M \times I$. (This term will also be used for a (closed) PL map $F : X \times I \to M$ whose restriction to each $X \times \{t\}$, $t \in I$, is an embedding.) An *isotopy* of M is a level-preserving PL homeomorphism $H : M \times I \to M \times I$ such that $H_0 = \mathrm{id}$. An isotopy F of X in M is *ambient* if there is an isotopy H of M making the following diagram commute.



We compose isotopies F and G of M by "stacking":

$$F \circ G(x, t) = \begin{cases} F(x, 2t), & \text{if } 0 \leqslant t \leqslant 1/2; \\ G(F(x, 1), 2t - 1), & \text{if } 1/2 \leqslant t \leqslant 1. \end{cases}$$

PROPOSITION 3.11 (Alexander Isotopy). *If $h_0, h_1 : B^n \to B^n$ are PL homeomorphisms that agree on S^{n-1}, then h_0 and h_1 are ambient isotopic by an isotopy that fixes S^{n-1}.*

PROOF. As $B^n \cong v * S^{n-1}$, use Proposition 2.3 to get

$$B^n \times [-1, 1] \cong v * \big(S^{n-1} \times [-1, 1] \cup B^n \times \{-1, 1\}\big).$$

Define

$$H : S^{n-1} \times [-1, 1] \cup B^n \times \{-1, 1\} \to S^{n-1} \times [-1, 1] \cup B^n \times \{-1, 1\}$$

by $H|S^{n-1} \times [-1, 1] \cup B^n \times \{-1\} = \mathrm{id}$ and $H|B^n \times \{1\} = h_1 \circ h_0^{-1}$. Extend linearly over the cone to get $H : B^n \times [-1, 1] \to B^n \times [-1, 1]$. (The Alexander Isotopy is the isotopy $Hh_0|B^n \times I : B^n \times I \to B^n \times I$.) □

PROPOSITION 3.12. *If X is collared in M, then any isotopy of X extends to an isotopy of M supported on a collar of X in M.*

The proof of this proposition as well as the following corollary to 3.11 and 3.12 are left as exercises.

COROLLARY 3.13. *If C is a cell complex and $f : C \to C$ is a homeomorphism that carries each cell of C onto itself, then f is ambient isotopic to the identity.*

THEOREM 3.14 (Regular Neighborhood Theorem). *Suppose X is a subpolyhedron in the interior of a PL manifold M and N_1 and N_2 are regular neighborhoods of X in* $\operatorname{int} M$. *Then there is an isotopy of M, fixed on X and outside an arbitrary neighborhood of $N_1 \cup N_2$ taking N_1 to N_2.*

PROOF. Let $N_0 \subseteq \operatorname{int} N_1 \cap \operatorname{int} N_2$ be a regular neighborhood of X. Then $C\ell(N_i - N_0) \cong \partial N_0 \times I$ for $i = 1, 2$. For a given neighborhood U of $N_1 \cup N_2$, choose regular neighborhoods N_i^+ of $C\ell(N_i - N_0)$ in $U - X$, $i = 1, 2$. Then there is a PL homeomorphism $h_i : N_i^+ \to \partial N_0 \times [0, 3]$ such that $h_i(C\ell(N_i - N_0), \partial N_0, \partial N_i) = (\partial N_0 \times [1, 2], \partial N_0 \times \{1\}, \partial N_0 \times \{2\}$. There is an obvious ambient isotopy of $\partial N_0 \times [0, 3]$, fixing $\partial N_0 \times \{0, 3\}$, taking $\partial N_0 \times \{2\}$ to $\partial N_0 \times \{1\}$. An appropriate composition does the job. □

Collapsing and shelling. Theorem 3.5 leads toward another important characterization of regular neighborhoods, because of the very special way in which a simplicial mapping cylinder deforms to its range. If $X \supset Y$ are polyhedra such that, for some $n \geqslant 0$,

$$\big(C\ell(X - Y), C\ell(X - Y) \cap Y\big) \cong \big(B^{n-1} \times I, B^{n-1} \times \{0\}\big),$$

then we say that there is an *elementary collapse* from X to Y, $X \searrow_e Y$. We say that X *collapses to* Y, $X \searrow Y$, if there is a sequence of elementary collapses

$$X = X_0 \searrow_e X_1 \searrow_e X_2 \searrow_e \cdots \searrow_e X_k = Y.$$

If $X \searrow Y$, then Y *expands* to X, $Y \nearrow X$. A (compact) polyhedron X is *collapsible*, $X \searrow 0$, if X collapses to a point.

If $M \supset Q$ are PL n-manifolds and $M \overset{e}{\searrow} Q$, then we call the elementary collapse an *elementary shelling*. If we set $(B^n, B^{n-1}) = (C\ell(M - Q), C\ell(M - Q) \cap Q)$, then $B^{n-1} \subseteq \partial Q$, and, hence, there is a homeomorphism $h : M \to Q$, fixed outside any preassigned neighborhood of $\operatorname{int} B^{n-1}$ in ∂Q. We say that M *shells* to Q if there is a sequence of elementary shellings starting with M and ending with Q.

PROPOSITION 3.15. *If $f : K \to L$ is a simplicial map with K finite, then $|C_f| \searrow |L|$.*

PROOF. A quick way to see this is to apply a result of M.H.A. Newman (see [66, Chapter 7, Lemma 46], also [15, Proposition 9.1]), which says that if σ and τ are simplexes, $\dim \sigma = n$, and if $f : \sigma \to \tau$ is a linear surjection, then $(|C_f|, \sigma) \cong (B^n \times I, B^n \times \{0\})$. (The proof of this assertion is not as immediate as one might like.) One then proceeds by induction downward through the skeleta of K.

It is also possible to prove this directly, using induction and the fact that, if X is a compact polyhedron, $C(C(X)) \searrow C(X)$.

Clearly, if $X \searrow Y$, then X deformation retracts to Y, but the converse may fail be true in a very strong sense. Polyhedra X and Y are *simple homotopy equivalent* if there is a sequence $X = X_0 \searrow X_1 \nearrow X_2 \searrow X_3 \nearrow \cdots \searrow Y_k = Y$. In particular, if $f : K \to L$ is a simplicial map (of finite complexes), which is also a homotopy equivalence from $|K|$ to $|L|$, then $|C_f|$ deformation retracts to $|K|$, but the equivalence may not be simple. There is an obstruction $\tau_f \in \mathrm{Wh}(\pi_1(|K|))$, the Whitehead group of the fundamental group of $|K|$: for a homotopy equivalence $f : |K| \to |L|$, $\tau_f = 0$ if, and only if, the inclusion of $|K|$ in $|C_f|$ is a simple homotopy equivalence. We refer the reader to [17] for a comprehensive treatment of this topic.

We state the collapsibility criteria for regular neighborhoods. They depend upon the fact that if $X \searrow Y$, then a regular neighborhood of X shells to a regular neighborhood of Y. Complete proofs may be found in [50] and [66]. □

THEOREM 3.16. *Suppose X is a compact polyhedron in the interior of a PL manifold M. A polyhedral neighborhood N of X in* $\operatorname{int} M$ *is a regular neighborhood of X if and only if*

(i) *N is a compact manifold with boundary,*

(ii) $N \searrow X$.

COROLLARY 3.17. *If $X \searrow 0$, then a regular neighborhood of X in a PL manifold is a ball.*

There are analogues for these results in the case of noncompact polyhedra and "proper" maps. The reader is referred to [51] and [53] for more details.

Regular neighborhoods of pairs. The Simplicial and Regular Neighborhood Theorems can be generalized to the "proper" inclusion of polyhedral pairs $(Y, Y_0) \subseteq (X, X_0)$, meaning $Y \cap X_0 = Y_0$. The simplicial model is constructed as before: Let $(K, K_0) > (L, L_0)$ be triangulations of $(X, X_0) \supset (Y, Y_0)$, with $L \lhd K$. Then $L_0 \lhd K_0$ and polyhedra $N_0 \subseteq N$

of the derived neighborhoods $N(L_0, K'_0) < N(L, K')$ are regular neighborhoods of Y_0 in X_0 and Y in X, respectively. Call (N, N_0) a *regular neighborhood of the pair* (Y, Y_0) in (X, X_0).

We will mostly be interested in the case in which X and X_0 are PL manifolds. Suppose Q is a q-dimensional submanifold of a PL n-manifold M. We say that Q is *proper* in M if $Q \cap \partial M = \partial Q$, and, if $Q \subseteq M$ is proper, we call the pair (M, Q) an (n, q)-*manifold pair*. A proper ball pair (B^n, B^q) is *unknotted* if $(B^n, B^q) \cong (J^n, J^q \times \{0\})$, where $J = [-1, 1]$. Similarly, a sphere pair (S^n, S^q) is *unknotted* if $(S^n, S^q) \cong (\partial J^{n+1}, \partial J^{q+1} \times \{0\})$. A manifold pair (M, Q) is *locally flat at* $x \in Q$ if there is a triangulation $K > L$ of $M \supset Q$, containing x as a vertex, such that the pair $(\mathrm{st}(v, K), \mathrm{st}(v, L))$ is an unknotted ball pair. (In the case that $Q \subseteq M$ is not proper and $x \in \partial Q - \partial M$, require instead that $(\mathrm{st}(v, K), \mathrm{st}(v, L)) \cong (J^n, J^{q-1} \times [0, 1] \times \{0\})$.) (M, Q) is a *locally flat manifold pair* if it is locally flat at every point. It is an exercise to see that (M, Q) is a locally flat manifold pair if there is a triangulation $K > L$ of $M \supset Q$ such that $(\mathrm{st}(v, K), \mathrm{st}(v, L))$ is an unknotted ball pair for each vertex v of L.

We state the Regular Neighborhood Theorem for Pairs. The proof follows that of that of the Regular Neighborhood Theorem with the obvious changes.

THEOREM 3.18 (Regular Neighborhood Theorem for Pairs). *Suppose (X, Y) is a polyhedral pair in a locally flat manifold pair (M, Q), with $X \cap Q = Y$, and suppose $(N_1, N_{1,0})$ and $(N_2, N_{2,0})$ are regular neighborhoods of (X, Y) in (M, Q). Then there is an isotopy H of (M, Q), fixed on X and outside a neighborhood of $N_1 \cup N_2$ with $H_1(N_1, N_{1,0}) = (N_2, N_{2,0})$.*

If (M, Q) is a locally flat manifold pair, then the pair $(\partial M, \partial Q)$ is locally collared *as pairs* in (M, Q). That is, if $x \in \partial Q$, then x has a neighborhood pair $(X, Y) \subseteq (M, Q)$ and $(X_0, Y_0) \subseteq (\partial M, \partial Q)$ such that $(X, X_0, Y, Y_0) \cong (X_0 \times [0, 1], X_0 \times \{0\}, Y_0 \times [0, 1], Y_0 \times \{0\})$. The proof of Theorem 2.4 generalizes immediately to provide a collaring theorem for pairs.

THEOREM 3.19. *If $(Y, Y_0) \subseteq (X, X_0)$ is a proper inclusion of polyhedral pairs and (Y, Y_0) is locally collared in (X, X_0) at each point of Y_0, then (Y, Y_0) is collared in (X, X_0).*

COROLLARY 3.20. *If (M, Q) is a locally flat manifold pair, then $(\partial M, \partial Q)$ is collared in (M, Q).*

One may define collapsing and shelling for pairs. For example, $(X, X_0) \searrow (Y, Y_0)$ means that $X_0 \cap Y = Y_0$, $X \searrow Y$, $X_0 \searrow Y_0$, and the collapse preserves X_0. For example, if $X \overset{e}{\searrow} Y$ so that $X = Y \cup B$, where B is a cell meeting Y in a face C, then $B \cap X_0$ must be a cell meeting Y_0 in a face that lies in C. In particular one can arrange that $X \searrow X_0 \cup Y \searrow Y \searrow Y_0$.

THEOREM 3.21. *If $(Y, Y_0) \subseteq (X, X_0) \subseteq (M, Q)$ are proper inclusions of pairs and (M, Q) is a locally flat manifold pair, and if $(X, X_0) \searrow (Y, Y_0)$, then a regular neighborhood pair of (X, X_0) in (M, Q) shells to one of (Y, Y_0).*

COROLLARY 3.22. *If $(X, X_0) \subseteq (M, Q)$ is a proper inclusion, where (M, Q) is a locally flat manifold pair, and if $(X, X_0) \searrow 0$, then a regular neighborhood pair of (X, X_0) in (M, Q) is an unknotted ball pair.*

Cellular moves. Two q-dimensional, locally flat submanifolds Q_1, Q_2 of a PL n-manifold M differ by a *cellular move* if there is a $(q+1)$-ball $B^{q+1} \subseteq \operatorname{int} M$ meeting Q_1 and Q_2 in complementary q-balls B_1^q and B_2^q, respectively, in ∂B^{q+1} such that $Q_1 \cap Q_2 = Q_1 - \operatorname{int} B_1^q = Q_2 - \operatorname{int} B_2^q$.

THEOREM 3.23. *If $Q_1, Q_2 \subseteq M$ differ by a cellular move across a $(q+1)$-ball B^{q+1}, then there is an isotopy H of M, fixed outside an arbitrary neighborhood of B^{q+1}, such that $H_1(Q_1) = Q_2$.*

PROOF. Using derived neighborhoods, we can get a regular neighborhood N of B^{q+1} in M such that if $N_i = N \cap Q_i$, then (N, N_i), $i = 1, 2$, is a regular neighborhood pairs of (B^{q+1}, B_i^q) in (M, Q_i). Since, by Corollary 3.22, each (N, N_i) is an unknotted ball pair there is a homeomorphism $h : (N, N_1) \to (N, N_2)$, fixed on the boundary. The Alexander Isotopy provides the isotopy H. □

COROLLARY 3.24. *A locally flat sphere pair (S^n, S^q) is unknotted iff S^q bounds a $(q+1)$-ball in S^n.*

PROOF. If $S^q = \partial B^{q+1}$, choose a triangulation $K > L$ of $S^n \supset B^{q+1}$ and a $(q+1)$-simplex σ of L such that $\sigma \cap S^q$ is a q-dimensional face of σ. Then $(\operatorname{st}(\sigma, K), \sigma)$ is an unknotted ball pair, and $\partial\sigma$ and S^q differ by a cellular move. □

Relative regular neighborhoods. If $Z \supset X \supset Y$ are polyhedra, then one can define a relative regular neighborhood of X mod Y in Z. The simplicial model is constructed much as above: Choose a triangulation $J > K > L$ of $Z \supset X \supset Y$, let J'' be a second derived subdivision of J mod K, and set $N(K - L, J'') = \{\sigma \in J'' \colon$ for some $\tau \in J''$, $\sigma < \tau$ and $\tau \cap |K| - |L| \neq \emptyset\}$. We recommend [16] for a complete treatment, including recognition and uniqueness theorems. As an example result from the theory we have the following.

THEOREM 3.25. *Suppose that $(B^q, \operatorname{int} B^q) \subseteq (M, \operatorname{int} M)$ and $(\operatorname{int} M, \operatorname{int} B^q)$ is a locally flat pair. If N is a regular neighborhood of B^q mod ∂B^q in M, then (N, B^q) is an unknotted ball pair.*

Structure of regular neighborhoods. We have commented on the fact that a regular neighborhood N of a polyhedron Y in a polyhedron X has the structure of a mapping cylinder of a mapping $\phi : \dot{N} \to Y$. In [16], Section 5, Cohen analyzes the fine structure of the mapping cylinder projection $\gamma : N \to Y$.

THEOREM 3.26 ([16]). *If N is a regular neighborhood of Y in X, then for each $y \in Y$, $\gamma^{-1}(y) \cong y * \phi^{-1}(y)$. Moreover, if (X, Y) is a locally unknotted (n, q)-manifold pair, then $\phi^{-1}(y) \cong S^{n-q-1} \times B^i$, where i is an integer depending on y.*

Suppose now that (M, Q) is a locally flat (n, q)-manifold pair. It is not generally true that we can get the integer i in Theorem 3.26 to be 0 for all $y \in Q$. Whenever that is possible the regular neighborhood N of Q in M has the structure of an $(n-q)$-disk bundle over Q. There are, however, examples [25,48] of locally flat PL embeddings without disk-bundle neighborhoods (although, they acquire disk-bundle neighborhoods after stabilizing the ambient manifold). Rourke and Sanderson show [49] that it is possible, however, to give N the structure of a block bundle. Given polyhedra E, F, and X, a PL mapping $\phi: E \to X$ is a *(PL) block bundle with fiber F* if there are PL cell complex structures $\mathcal{K}$ and $\mathcal{L}$ on E and X, respectively, such that $\phi: \mathcal{K} \to \mathcal{L}$ is cellular and for each cell $C \in \mathcal{L}$, $\phi^{-1}(C)$ is PL homeomorphic to $C \times F$. If $\phi: E \to X$ is a block bundle with fiber F, then the mapping cylinder retraction $\gamma: C_\phi \to X$ is also a block bundle with fiber the cone $x * F$, and for each cell $C \in \mathcal{L}$, $(\gamma^{-1}(C), C) \cong (C \times (x * F), C \times \{x\})$. If $F = S^{m-1}$ and C is a p-cell in $\mathcal{L}$, then $(\gamma^{-1}(C), C) \cong (J^{p+m}, J^p \times \{0\})$. A PL retraction $\gamma: E \to X$ satisfying this property is called an *m-block bundle* over X.

THEOREM 3.27 ([49]). *Suppose that (M, Q) is a locally flat (n, q)-manifold pair. Then a regular neighborhood N of Q in M has the structure of an $(n-q)$-block bundle over Q.*

PROOF. We only consider the case in which $\partial Q = \emptyset$. Let $K > L$ be a triangulation of $M \supset Q$ with $L \lhd K$, let K_1 be a first derived subdivision of K mod L, and let $N = |N(L, K_1)|$. Since (M, Q) is a locally flat, for any p-simplex $\sigma \in L$, $(\mathrm{lk}(\sigma, K_1), \mathrm{lk}(\sigma, L))$ is an unknotted $(n-p-1, q-p-1)$-sphere pair; hence,

$$(\mathrm{lk}(\sigma, K_1), \mathrm{lk}(\sigma, L)) \cong (S^{q-p-1} * S^{n-q-1}, S^{q-p-1}).$$

Let $K' > L'$ be a first derived subdivision of $K > L$ extending K_1, and let σ be a p-simplex of L. Let $\widetilde{K}_\sigma < K'$ and $\widetilde{L}_\sigma < L'$ denote the dual $(n-p-1)$- and $(q-p-1)$-spheres to σ in K' and L', respectively, and let $\widetilde{C}_\sigma = \hat{\sigma} * \widetilde{K}_\sigma$ and $\widetilde{D}_\sigma = \hat{\sigma} * L_\sigma$ denote the respective dual cells. Then

$$(\widetilde{K}_\sigma, \widetilde{L}_\sigma) \cong (\mathrm{lk}(\sigma, K_1), \mathrm{lk}(\sigma, L)) \cong (S^{q-p-1} * S^{n-q-1}, S^{q-p-1}),$$

so that

$$(\widetilde{C}_\sigma, \widetilde{D}_\sigma) \cong (J^{n-p}, J^{q-p} \times \{0\}).$$

These dual cell pairs fit together nicely to give the neighborhood N the structure of an $(n-q)$-block bundle over Q with respect to the dual cell structures on M and Q obtained from K and L. The mapping $\gamma: N \to Q$ is obtained by induction on the dual cells of L; it is not, in general, the same as the natural projection defined above. □

4. General position

General position is a process by which two polyhedra X and Y in a PL manifold M may be repositioned slightly in order to minimize the dimension of $X \cap Y$. It is also a process by which the dimension of the singularities of a PL map $f: X \to M$ may be minimized by a small adjustment of f. A combination of general position and join structure arguments form the underpinnings of nearly every result in PL topology. We start with definitions of "small adjustments".

Given metric spaces X and M and $\varepsilon > 0$ (ε may be a continuous function of X), An *ε-homotopy* (*isotopy*) of X in M is a homotopy (isotopy) $F: X \times I \to M$ such that diam $F(x \times I) < \varepsilon$ for every $x \in X$. An *ε-isotopy* of M is an isotopy H of M that is also an ε-homotopy. If $X, Y \subseteq M$, then an *ε-push* of X in M, rel Y, is an ε-isotopy of M that is fixed on Y and outside the ε-neighborhood of X.

Suppose $f: X \to M$ is a (continuous) function. The *singular set* of f, is the subset $S(f) = C\ell\{x \in X\colon f^{-1}f(x) \neq x\}$. If X and M are polyhedra and f is PL, then f is *non-degenerate* if dim $f^{-1}(y) \leqslant 0$ for each $y \in M$. If f is a PL map, and $f^{-1}(C)$ is compact for every compact subset C of M, then $S(f)$ is a subpolyhedron of X.

Let us start with a (countable) discrete set S of points in $\mathbb{R}^n$. We say that S is in *general position* if every subset $\{v_0, v_1, \ldots, v_p\}$ of S spans a p-simplex, whenever $p \leqslant n$. Since the set of all hyperplanes of $\mathbb{R}^n$ of dimension $< n$ spanned by points of S is nowhere dense, it is clear that if $\varepsilon: S \to (0, \infty)$ is arbitrary, then there is an isotopy H of $\mathbb{R}^n$, fixed outside an ε-neighborhood of S such that $H_1(S)$ is in general position and diam $H(v \times I) < \varepsilon(v)$ for all $v \in S$. Moreover, if S_0 is a subset of S that is already in general position, then we can require that H fixes S_0 as well. We can also approximate any map $f: S \to \mathbb{R}^n$ by map g such that $g(S)$ is in general position, insisting that $g|S_0 = f|S_0$ if $f(S_0)$ is already in general position.

General position properties devolve from the following elementary fact from linear algebra.

PROPOSITION 4.1. *Suppose that E_1, E_2 and E_0 are hyperplanes in $\mathbb{R}^n$ of dimensions p, q and r, respectively, spanned by $\{u_0, u_1, \ldots, u_p\}$, $\{v_0, v_1, \ldots, v_q\}$, and $\{w_0, w_1, \ldots, w_r\}$ with $u_i = v_i = w_i$ for $0 \leqslant i \leqslant r$ and $u_i \neq v_j$ for $i, j > r$. If the set $S = \{u_0, u_1, \ldots, u_p, v_{r+1}, v_{r+2}, \ldots, v_q\}$ is in general position, then* $\dim((E_1 - E_0) \cap (E_2 - E_0)) \leqslant p + q - n$.

As usual, we interpret $\dim(A \cap B) < 0$ to mean that $A \cap B = \emptyset$. Proposition 4.1 motivates the definition of general position for polyhedra X and Y embedded in a PL manifold M. If dim $X = p$, dim $Y = q$, and dim $M = n$, we say that X and Y are in *general position* in M if $\dim(X \cap Y) \leqslant p + q - n$.

THEOREM 4.2. *Suppose that $X \supset X_0$ and Y are polyhedra in the interior of a PL n-manifold M with* $\dim(X - X_0) = p$ *and* $\dim Y = q$ *and $\varepsilon: M \to (0, \infty)$ is continuous. Then there is an ε-push H of X in M,* rel X_0, *such that* $\dim[H_1(X - X_0) \cap Y] \leqslant p + q - n$.

PROOF. Let $J > K, K_0$ be a triangulation of $M \supset X, X_0$ with $K_0 \lhd K$. Let v be a vertex of $K - K_0$. Let $g: \mathrm{lk}(v, J) \to S^{n-1}$ be a PL homeomorphism that is linear on each simplex of $\mathrm{Lk}(v, J)$. Extend g conewise to a PL homeomorphism $h: \mathrm{st}(v, J) \to B^n$ such that $h(v) = 0$. Apply Proposition 4.1 to get a point $x \in \mathring{B}^n$ such that $\dim(((x * \tau) - \tau) \cap h(Y)) \leqslant p + q - n$ for every simplex $\tau = h(\sigma), \sigma \in \mathrm{Lk}(v, K)$. Let F be an isotopy of B^n, fixed on S^{n-1}, with F_1 the conewise extension of $\mathrm{id}_{S^{n-1}}$ that takes x to 0. Let F^v be the isotopy of M, fixed outside $\mathrm{st}(v, J)$, obtained by conjugating F with h. By choosing x sufficiently close to $0 \in B^n$, we can assume that F^v is a δ-push of Y in M, rel $(M - \mathring{\mathrm{st}}(v, J))$. Then $\dim((\mathrm{st}(v, K) - \mathrm{lk}(v, K)) \cap F_1^v(Y)) \leqslant p + q - n$, and for any $\delta > 0$.

Assume now that K is a derived subdivision of a triangulation of X so that the vertices of K can be partitioned: $K^{(0)} = V_0 \cup V_1 \cup \cdots \cup V_k$, $k = \dim K$, where

$$\mathrm{st}(v, K) \cap \mathrm{st}(w, K) \subseteq \mathrm{lk}(v, K) \cap \mathrm{lk}(w, K)$$

when $v, w \in V_i$, $v \neq w$. (See the proof of Theorem 2.4.) For $0 \leqslant i \leqslant k$, define an isotopy F^i of M by

$$F^i = F^v \quad \text{on } \mathrm{st}(v, J),\ v \in V_i,$$

and

$$F^i = \mathrm{id} \quad \text{on } M - \bigcup_{v \in V_i} \mathring{\mathrm{st}}(v, J).$$

We can easily make F^i an $\frac{\varepsilon}{2(k+1)}$-push of X in M, rel $(M - \bigcup_{v \in V_i} \mathring{\mathrm{st}}(v, J))$. If we construct the F^i's inductively we can ensure that the composition $G = F^0 \circ \cdots \circ F^k$ is an $\frac{\varepsilon}{2}$-push of X in M, rel X_0, and that $\dim((X - X_0) \cap G_1(Y)) \leqslant p + q - n$. The inverse H of G is then an ε-push of X in M, rel X_0, such that $\dim(H_1(X - X_0) \cap Y) \leqslant p + q - n$. (The inverse of an ε-push H of X is only a 2ε-push of $H(X)$.) □

A similar type of argument can be used to prove a general position theorem for mappings.

THEOREM 4.3. *Suppose $X \supset X_0$ are polyhedra with $\dim(X - X_0) = p$, M is a PL n-manifold, $p \leqslant n$, and $f : X \to M$ is a continuous map with $f|X_0$ PL and nondegenerate on some triangulation of X_0. Then for every continuous $\varepsilon : X \to (0, \infty)$ there is an ε-homotopy, rel X_0, of f to $f' : X \to M$ such that $\dim(S_f - X_0) \leqslant 2p - n$. Moreover, if $X_1 \subseteq X$ and $\dim(X_1 - X_0) = q$, then we can arrange to have $\dim((S_f \cap (X_1 - X_0)) \leqslant p + q - n$.*

A mapping satisfying this last condition is said to be *in general position with respect to* X_1 rel X_0.

COROLLARY 4.4. *Suppose $X \supset X_0$ is a p-dimensional polyhedron, $f : X \to M$ is a continuous mapping of X into a PL n-manifold M, $2p + 1 \leqslant m$, such that $f|X_0$ is a PL embedding, and if $\varepsilon : X \to (0, \infty)$ is continuous, then f is ε-homotopic, rel X_0, to a PL embedding.*

This is the best one can expect in such full generality. There is a p-dimensional polyhedron X, namely the p-skeleton of a $(2p + 2)$-simplex, that does not embed in $\mathbb{R}^{2p}$ [20].

Shapiro [52] has developed an obstruction theory for embedding p-dimensional polyhedra in $\mathbb{R}^{2p}$.

General position and regular neighborhood theory can be used to establish an unknotting theorem for sphere pairs.

THEOREM 4.5. *A sphere pair* (S^n, S^q) *is unknotted, if*
(i) $q = 1$ *and* $n = 4$, *or*
(ii) $n \geqslant 2q + 1$ *and* $n \geqslant 5$.

COROLLARY 4.6. *An* (n, q)*-manifold pair* (M, Q) *is locally flat provided* $q = 1$, $n \geqslant 1$, *or* $q = 2$, $n \geqslant 5$, *or* $q > 2$, $n \geqslant 2q$.

PROOF ([50]). (i) If $n \geqslant 5$, then general position gives an embedding of the cone on S^1, so that S^1 is unknotted by 3.24.

If $n = 4$, then there is a point $x \in \mathbb{R}^4$ such that x and S^1 are joinable: Let $V = \bigcup\{E(u, v)\colon u, v \in S^1\}$, where $E(u, v)$ is the line determined by u and v. V is a finite union of hyperplanes, each of dimension at most 3 in $\mathbb{R}^4$. Hence, if $x \notin V$, then $x * S^1$ is the cone on S^1. Thus S^1 bounds a 2-ball in $\mathbb{R}^4$.

(ii) Assume as above that $S^q \subseteq \mathbb{R}^n$. By induction, using Corollary 4.6, we may assume that (S^n, S^q) is locally flat. Since $2q \leqslant n - 1$, we may assume that the restriction of the projection $\pi\colon \mathbb{R}^n \to \mathbb{R}^{n-1}$ to S^q has a singular set consisting of double points $\{a_1, b_1', \ldots, a_r, b_r'\}$, where a_i lies "above" b_i'. Choose a point x near infinity "above" S^q, and let $f\colon x * S^q \to \mathbb{R}^n$ be the natural linear extension to the cone $x * S^q$, so that the singularities of f lie in $\bigcup_{i=1}^r x * \{a_i, b_i\}$, where b_i is close to b_i'. Since $q \geqslant 2$, there is a PL q-cell B in $S^q - \{b_1, \ldots, b_r\}$ containing $\{a_1, \ldots, a_r\}$. Then $f|x * \partial B$ is an embedding as is $f|x * (S^q - \operatorname{int} B)$. The $(q + 1)$-ball $f(x * B)$ provides a cellular move from S^q to $\partial f(x * (S^q - \operatorname{int} B))$. But $\partial f(x * (S^q - \operatorname{int} B))$ is unknotted, by 3.24. □

THEOREM 4.7 ([6]). *Suppose* $X \supset X_0$ *is a* p*-dimensional polyhedron,* M *is a PL* n*-manifold,* $2p + 2 \leqslant n$, *and* $f, g\colon X \to M$ *are PL embeddings such that* $f|X_0 = g|X_0$ *and* $f \simeq g$, rel X_0. *Then* f *and* g *are ambient isotopic,* rel X_0, *by an isotopy supported on an arbitrary neighborhood of the image of a homotopy of* f *to* g.

PROOF. Let $K > K_0$ be a triangulation of $X \supset X_0$ and assume, inductively, that $f|K_0 \cup K^{(p-1)} = g|K_0 \cup K^{(p-1)}$. As the isotopy will be constructed by moving across balls with disjoint interiors, we assume further that $f \simeq g$, rel $|K_0| \cup |K^{(p-1)}|$. Let $Z = X \times I$ mod $(|K_0| \cup |K^{(p-1)}|)$, and let $F\colon Z \to M$ be a relative homotopy from f to g. Assume F is PL and in general position, so that $\dim S(F) \leqslant 2(p + 1) - n \leqslant 0$ and $S(F)$ consists of double points $\{a_i, b_i\}$ lying in the interiors of cells $\sigma \times I$ mod $\partial\sigma$, where σ is a p-simplex of K. For each $a_i \in \operatorname{int}\sigma \times (0, 1)$, get a PL arc A_i in $\operatorname{int}\sigma \times [0, 1)$ joining a_i to a point $c_i \in \operatorname{int}\sigma \times \{0\}$, chosen so that the A_i's are disjoint and contain none of the b_j's. Get a regular neighborhood pair $(D_i, D_{i,0})$ of (A_i, c_i) in $(\operatorname{int}\sigma \times [0, 1), \operatorname{int}\sigma \times \{0\})$, chosen so that the D_i's are disjoint (and contain none of the b_j's). Let C_i be the face of ∂D_i complementary to $D_{i,0}$. Then there is a cellular move across D_i taking $\sigma_i \times \{0\}$ to $(\sigma_i \times \{0\}) - \operatorname{int} D_{i,0}$. The net effect of these moves is to get a homotopy of F to an embedding.

Assume now that we have an embedding $F : Z \to M$. For p-simplexes $\sigma \in K$, choose relative regular neighborhoods N_σ of $F(\sigma \times I)$ mod $F(\partial\sigma)$ so that $N_\sigma \cap F(Z) = F(\sigma \times I)$ and the N_σ's have disjoint interiors. Then $(N_\sigma, F(\sigma \times \{i\}))$, $i = 0, 1$, is an unknotted ball pair. Hence, there is an isotopy H of M, fixed outside the union of the N_σ's, such that on N_σ, $H_1 \circ f|\sigma = g|\sigma$. □

5. Embeddings, engulfing

In this section we address the following question, which arises naturally from Corollary 4.4. Suppose X is a p-dimensional polyhedron and M is a PL n-manifold. When is a map $f : X \to M$ homotopic to a PL embedding? The first theorem takes a small but important step in reducing the codimension restriction of Corollary 4.4.

THEOREM 5.1. *Suppose that Q is a connected PL q-manifold, M is a properly simply connected PL n-manifold, and $n \geqslant 2q \neq 4$. Then every closed mapping $f : Q \to M$ is homotopic to a PL embedding.*

PROOF. To say that M is *properly* simply connected means that M is simply connected and *simply connected at infinity*. That is, for every compact set C in M there is a compact set $D \supseteq C$ such that any loop in $M - D$ is null-homotopic in $M - C$. We consider only the case $n \geqslant 6$. The proof exploits the now famous "Whitney Trick" [64,62]. Given $f : Q \to M$, use general position to get a PL mapping $g : Q \to \operatorname{int} M$ homotopic to f such that $S(g)$ is a closed set consisting only of "double points": $S(g) = \{a_1, b_1, a_2, b_2, \ldots\} \subseteq \operatorname{int} Q$, where the indicated points are distinct, $g(a_i) = g(b_i)$, $i = 1, 2, \ldots$, and $g(a_i) \neq g(a_j)$ if $i \neq j$. Since $q \geqslant 3$, we can get a closed family of mutually exclusive PL arcs $A_1, A_2, \ldots$ joining a_i to b_i, respectively. The images $g(A_i)$ are PL simple closed curves in $\operatorname{int} M$. Since M is properly simply connected and $n \geqslant 6$, we can use general position to get a closed family of mutually exclusive PL 2-cells $D_1, D_2, \ldots$ in $\operatorname{int} M$ such that $\partial D_i = g(A_i)$. Using suitable triangulations we can get mutually exclusive regular neighborhoods N_i of D_i in $\operatorname{int} M$ such that $V_i = g^{-1}(N_i)$ is a regular neighborhood of A_i in $\operatorname{int} Q$, $i = 1, 2, \ldots$. By Corollary 3.17 N_i and V_i are PL balls of dimensions n and q, respectively. Using the cone structures on N_i and V_i, we can redefine $g|V_i$ to get an embedding $h_i : V_i \to N_i$, agreeing with g on ∂V_i, and homotopic to $g|V_i$ rel ∂V_i. Then $g \simeq h$, where $h|V_i = h_i$ and $h|Q - \bigcup_i V_i = g|Q - \bigcup_i V_i$. □

Generalizations of the Whitney Trick may be used to reduce the codimension, $n - q$, provided compensating assumptions are made on the connectivity of Q and M. One approach uses engulfing techniques, introduced by Stallings [56] and Zeeman [66], which have proved useful in other contexts as well.

Engulfing. The engulfing problem: Given a closed set Y (polyhedron) and a compact set C in a PL manifold M, with $C \subseteq \operatorname{int} M$, and an arbitrary neighborhood U of Y in M, find an ambient isotopy H of M, fixed on $Y \cup \partial M$ and outside a compact set, such that $H_1(U) \supset C$. If such an isotopy of M exists, we say that C *can be engulfed from* Y. Obvious homotopy conditions must be met, but they are not sufficient in general to find H. One need only look at the Whitehead link as C in the torus $S^1 \times B^2$, with $Y = pt$. (See, e.g., [66, Chapter 7].)

THEOREM 5.2. *Suppose Y is a compact polyhedron of dimension $\leqslant n-3$ in a PL n-manifold M, such that (M, Y) is k-connected. A compact, k-dimensional polyhedron X in* $\operatorname{int} M$ *can be engulfed from Y provided*

(i) $n \geqslant 6$ *and* $n-k \geqslant 3$, *or*
(ii) $n = 4$ *or* 5 *and* $k = 1$, *or*
(iii) $n = 5$, $k = 2$.

PROOF. The proof uses the collapsing techniques of [56] and [66]. We shall first give an argument for (i) in the case $n-k \geqslant 4$, deferring the case $n-k = 3$ of (i) and (iii). We leave the proof of (ii) as an exercise. An elegant alternative proof of Theorem 5.2, using handle theory, may be found in [50].

The proof uses the fact that a simplicial mapping cylinder collapses to its range. Suppose $A \supset B$ are polyhedra and $A \searrow B$. For a subset C of A, define the *trail* of C, $\operatorname{tr}(C) \supset C$, under the collapse as follows. Let $A = A_0 \overset{e}{\searrow} A_1 \overset{e}{\searrow} \cdots \overset{e}{\searrow} A_r = B$ be a sequence of elementary collapses, so that

$$\big(\mathcal{C}\ell(A_{i-1} - A_i), \mathcal{C}\ell(A_{i-1} - A_i) \cap A_i\big) \overset{h_i}{\cong} \big(B^{m-1} \times I, B^{m-1} \times \{0\}\big).$$

Suppose $\operatorname{tr}_i(C) = \operatorname{tr}(C) \cap \mathcal{C}\ell(A - A_i)$ has been defined for $0 \leqslant i < k$ (where $\operatorname{tr}_0(C) = \emptyset$). Let

$$D = h_k\big(\operatorname{tr}_{k-1}(C) \cup C\big) \cap \big(\mathcal{C}\ell(A_{k-1} - A_k)\big) \subseteq B^{m-1} \times I,$$

and let $E = \{(x,t) \in B^{m-1} \times I\colon (x,s) \in D, \text{ for some } s \geqslant t\}$. Define $\operatorname{tr}_k(C) = \operatorname{tr}_{k-1}(C) \cup h_k^{-1}(E)$. Finally, define $\operatorname{tr}(C) = \operatorname{tr}_r(C) \cup C$. If C is a polyhedron of dimension p in A, then elementary arguments show that

(a) $A \searrow B \cup \operatorname{tr}(C) \searrow B$, and
(b) $\dim \operatorname{tr}(C) \leqslant p+1$. □

Suppose now that $Y, X \subseteq M$, as in (i), with $n-k \geqslant 4$. We shall actually prove the stronger

ASSERTION 5.3. *There is a polyhedron $Q \subseteq M$ such that $X \subseteq Q$, $Q \searrow Y$, and* $\dim(Q - Y) \leqslant k+1$.

Given the assertion, one may apply the Regular Neighborhood Theorem to obtain the desired isotopy.

PROOF OF ASSERTION 5.3. Fix $k (\leqslant n-4)$, and suppose inductively that, for $0 \leqslant i \leqslant k$, we have the following:

(1) a polyhedron $Q \supset X \cup Y$ in M with $\dim Q \leqslant n-3$, such that
(2) $Q \searrow Y \cup P$, where
(3) $\dim P \leqslant k-i$.
Start the induction at $i=0$ with $Q = X \cup Y$ and $P = X$.

Since (M, Y) is k-connected, there is a homotopy of the inclusion of P in M, rel $P \cap Y$, to a map $f: P \to Y$, which we may assume to be PL. Choose triangulations K and L of P and Y, respectively, such that $H = K \cap L$ triangulates $P \cap Y$ and $f: K \to L$ is simplicial. Let $Z = |C_f \operatorname{rel} H|$. Then $\dim(Z-Y) \leqslant k-i+1$ and the homotopy provides a map $F: Z \to M$ such that $F|P \cup Y = \mathrm{id}$. We may assume that F is in general position (with respect to Y) so that $\dim S(F) \leqslant (n-3)+(k-i+1)-n \leqslant k-i-2$ (Theorem 4.3). Let $T = \mathrm{tr}(S(F))$ under the collapse $Z \searrow Y$. Then $\dim T \leqslant k-i-1$, and $Z \searrow Y \cup T \searrow Y$; hence, $F(Z) \searrow Y \cup F(T)$. Let $R = \mathrm{tr}(F(Z) \cap Q)$ under the collapse $Q \searrow Y \cup P$. Then $\dim R \leqslant k-i-1$, and $Q \searrow Y \cup P \cup R \searrow Y \cup P$.

Set $Q_1 = Q \cup S(Z)$ and $P_1 = F(T) \cup R$. Then $Q_1 \searrow Y \cup F(Z) \cup R \searrow Y \cup F(T) \cup R = Y \cup P_1$, and $\dim P_1 \leqslant k-i-1$.

When $i = k+1$, the process stops, since the set $P = \emptyset$. □

The inductive argument given above does not work in the case $n-k=3$. (Check the dimension of the polyhedron T in the proof.) To argue this case we shall use Zeeman's Piping Lemma, which we paraphrase next. A proof may be found in [66, Chapter 7, Lemma 48].

LEMMA 5.4 (Piping Lemma [66]). *Suppose M is a PL n-manifold, K is a finite complex of dimension $k \leqslant n-3$, $f: K \to L$, $\dim L \leqslant n-3$, is a simplicial mapping that restricts to an embedding on a subcomplex $H < K$, $Z = |C_f \operatorname{rel} H|$, $Z_0 = |C_{f|K^{(k-1)}} \operatorname{rel} H|$, and $F: Z \to M$ is a PL mapping that is in general position with respect to Z_0. Then F is homotopic* rel $|K| \cup Z_0$ *to a PL mapping $G: Z \to M$ such that*
(a) $Z \searrow Z_1 \searrow |L|$,
(b) $S(G) \cup Z_0 \subseteq Z_1$,
(c) $\dim(Z_1 - |L|) \leqslant k-1$, *and*
(d) $\dim \mathcal{C}\ell(Z_1 - |L|) \cap Z_0) \leqslant k-2$.

We indicate the proof of Assertion 5.3 when $k = n-3$, $n \geqslant 5$. The inclusion of X in M is homotopic rel $X \cap Y$ to a mapping $f: X \to Y$, which we may assume to be PL. Let K, L, H triangulate X, Y, $X \cap Y$ so that $f: K \to L$ is simplicial and $f|H = \mathrm{id}$. Let $Z = |C_f \operatorname{rel} H|$, and let $F: Z \to M$ be a PL mapping with $F|X \cup Y = \mathrm{id}$, guaranteed by the connectivity, in general position with respect to $Z_0 = |C_{f|K^{(n-4)}} \operatorname{rel} H|$. By Lemma 5.4 F is homotopic, rel $|K| \cup Z_0$, to a PL mapping $G: Z \to M$ satisfying (a), (b). and (c) of 5.4. Then $G(Z) \searrow G(Z_1)$. In the proof of Assertion 5.3, set $Q = G(Z)$ and $P = \mathcal{C}\ell(G(Z_1) - Y)$. Then $\dim Q \leqslant n-2$ and $\dim P \leqslant n-4$, and the inductive argument proceeds without a problem.

Generalizations of the Whitney Trick for eliminating double point singularities of a mapping $f: Q^q \to M^{2q}$, as in Theorem 5.1, can be obtained from the engulfing techniques just described. Irwin's embedding theorem, which we now state, can be thought of as the generalization to codimension 3 of the process of removing one pair of double points.

THEOREM 5.5 ([33]). *Suppose Q is a compact PL q-manifold, M is a PL n-manifold, $n - q \geqslant 3$, such that Q is $(2q - n)$-connected and M is $(2q - n + 1)$-connected. Then every map $f:(Q, \partial Q) \to (M, \partial M)$ for which $f|\partial Q:\partial Q \to \partial M$ is a PL embedding is homotopic* rel ∂Q *to a PL embedding.*

PROOF. Since general position works when $n \leqslant 5$, we assume that $n \geqslant 6$. By playing with the collar structures on ∂Q and ∂M, one may assume that $f(\operatorname{int} Q) \subseteq \operatorname{int} M$ and $f|N$ is a PL embedding for some collar neighborhood N of ∂Q in Q and that a general position approximation $g: Q \to M$ satisfies $g|N = f|N$, $S(g) \subseteq C\ell(Q - N)$, and $\dim S(g) \leqslant 2q - n$. We will find collapsible polyhedra $C \subseteq \operatorname{int} Q$ and $D \subseteq \operatorname{int} M$ such that $S(g) \subseteq C = g^{-1}(D)$. Once we have C and D, we can proceed as in the proof of Theorem 5.1: Get regular neighborhoods U of D in $\operatorname{int} M$ and V of C in $\operatorname{int} Q$ such that $g^{-1}(U) = V$. Then U and V are PL n- and q-balls, respectively, and $g|\partial V:\partial V \to \partial U$ is a PL embedding. We redefine g on V to get a PL embedding homotopic to g.

We shall assume, initially, that $n - q \leqslant 4$. We construct C and D by induction. Assume that $f: Q \to M$ is a PL mapping in general position with $S(f) \subseteq \operatorname{int} Q$ and $f(S(f)) \subseteq \operatorname{int} M$. Suppose, inductively, we have the following:

(a) polyhedra $C \subseteq Q$, $D \subseteq M$ such that
(b) $S(f) \subseteq C \searrow 0$, $f(C) \subseteq D \searrow 0$,
(c) $f^{-1}(D) = C \cup C_1$,
(d) $\dim C \leqslant q - 3$, $\dim D \leqslant q - 2$, and
(e) $\dim C_1 \leqslant q - 3 - i$, and $\dim(C_1 \cap C) \leqslant q - 4 - i$.

We start the induction at $i = 3$. Since $\dim S(g) \leqslant 2q - n \leqslant q - 4$, we can use the connectivity conditions and Assertion 5.3, with $Y = \text{pt}$, to get a collapsible polyhedron $C \subseteq Q$ with $\dim C \leqslant 2q - n + 1 \leqslant q - 3$. Apply the connectivity conditions and Assertion 5.3 again, with $Y = \text{pt}$, to get a collapsible polyhedron $D \supset f(C)$, with $\dim D \leqslant 2q - n + 2 \leqslant q - 2$. Use general position to get $\dim(f(Q - C) \cap D) \leqslant q + (q - 2) - n \leqslant q - 6$, and $\dim(C\ell(f(Q - C) \cap D) \cap D) \leqslant q - 7$. Set $C_1 = f^{-1}(C\ell(f(Q - C) \cap D))$; $\dim C_1 \leqslant q - 6$.

Suppose we are given (a)–(e), for $1 \leqslant i \leqslant k$, so that $\dim C_1 \leqslant q - 3 - k$, and $\dim(C_1 \cap C) \leqslant q - 4 - k$. Let $S = C_1 \cap C$, and let $T = \operatorname{tr} S$ under the collapse $C \searrow 0$; $\dim T \leqslant q - 3 - k$. Then $C \searrow T \searrow 0$. The connectivity conditions, together with the homotopy extension theorem, imply there is a homotopy, rel $C_1 \cap T$, of id_{C_1} to a mapping $g: C_1 \to T \subseteq C$. Assertion 5.3 then provides a polyhedron $A \supset C_1 \cup C$ such that $A \searrow C$ ($\searrow 0$), $\dim(A - C) \leqslant q - 2 - k$, and $C\ell(A - C) \cap C \leqslant q - 3 - k$.

Let $A_1 = C\ell(A - C)$ and let $B_1 = f(A_1)$. Then $\dim B_1 \leqslant q - 2 - k$ and $\dim B_1 \cap D \leqslant q - 3 - k$. Let $S_1 = B_1 \cap D$ and let $T_1 = \operatorname{tr} S_1$ under the collapse $D \searrow 0$. Then $\dim T_1 \leqslant q - 2 - k$ and $T_1 \searrow 0$. Repeat the argument above: use the connectivity conditions, together with the Homotopy Extension Theorem, to get a homotopy, rel $B_1 \cap T_1$, of id_{B_1} to a mapping $h: B_1 \to T_1 \subseteq D$. Assertion 5.3 then provides a polyhedron $P \supset B_1 \cup D$ such that $P \searrow D$ ($\searrow 0$), $\dim(P - D) \leqslant q - 1 - k$, and $C\ell(P - D) \cap D \leqslant q - 2 - k$. Use general position to get $\dim((P - D) \cap f(Q)) \leqslant (q - 1 - k) + q - n \leqslant q - 5 - k$. Set $P_1 = f^{-1}(C\ell(P - D))$. Then A and P_1 replace C and C_1 to complete the inductive step.

The case $n - q = 3$ requires Lemma 5.4 to get the induction going, very much as in the proof of Assertion 5.3. We shall leave the details to the reader. □

COROLLARY 5.6. *If Q is a compact k-connected q-manifold, $q - k \geqslant 3$, then Q embeds in $\mathbb{R}^{2q-k}$.*

COROLLARY 5.7. *If $f : S^{q-1} \to \partial M$ is a PL embedding of the $(q-1)$-sphere into the boundary of a $(q-1)$-connected n-manifold M, $n - q \geqslant 3$, then f extends to a PL embedding $\bar{f} : B^q \to M$.*

To get a generalization of the Whitney Trick analogous to the removal of *all* of the double point singularities of Theorem 5.1, one must impose a connectivity condition on the mapping f. Recall that a mapping $f : Q \to M$ is k-connected if $\pi_i(f) = \pi_i(M_f, Q) = 0$, for $0 \leqslant i \leqslant k$.

THEOREM 5.8 ([29,57,61]). *Suppose Q and M are PL manifolds of dimensions q and n, respectively, $n - q \geqslant 3$, and $f : (Q, \partial Q) \to (M, \partial M)$ is a $(2q - n + 1)$-connected map such that $f|\partial Q$ is a PL embedding. Then f is homotopic, rel ∂Q, to a PL embedding.*

This theorem was first proved by Hudson [29], with an extra connectivity hypothesis on Q, using a generalization of the techniques of the proof of Theorem 5.5. This condition later proved to be superfluous as a consequence of the argument in the proof of following theorem of Stallings [57]. We include Stallings' argument, since it has only appeared in preprint form.

THEOREM 5.9 ([57]). *Suppose X is a compact k-dimensional polyhedron, M is a PL manifold of dimension n, $n - k \geqslant 3$, and $f : X \to M$ is $(2k - n + 1)$-connected. Then there is a k-dimensional polyhedron $X_1 \subseteq M$ and a simple homotopy equivalence $f_1 : X \to X_1$ such that f_1 and f are homotopic as maps to M.*

PROOF ([57]). It is not difficult to see that a map $f : X \to M$ is i-connected if, and only if, any map $\alpha : (P, Q) \to (C_f, X)$ of a polyhedral pair (P, Q) into the mapping cylinder of f, with $\dim(P - Q) \leqslant i$, is homotopic, rel $\alpha|Q$, to a map into X.

Suppose that $f : X \to M$ is a $(2k - n + 1)$-connected map and that f is in general position, so that $\dim S_f \leqslant 2k - n$. Suppose, inductively, that we have a k-dimensional polyhedron Y, a simple homotopy equivalence $h : X \to Y$, and a PL map $g : Y \to M$ such that

(a) $gh = f$,

(b) $\dim S(g) \leqslant 2k - n - j$, for some j, $0 \leqslant j \leqslant 2k - n$.

Then g is $(2k - n + 1)$-connected. We start the induction by setting $Y = X$, $h = \mathrm{id}$, and $g = f$.

Set $S = S(g)$, $T = g(S(g))$, and let C be the mapping cylinder of $g|S : S \to T$ with projection $\gamma : C \to T$. Then C is a submapping cylinder of C_g and $\dim C \leqslant 2k - n - j + 1 \leqslant 2k - n + 1$. Our hypotheses imply that the inclusion $(C, S) \subseteq (C_g, Y)$ is homotopic, rel S, to a map of C into Y. Let $H : C \times I \to C_g$ be such a homotopy. That is, $H_0(y) = y$, for all $y \in C$, $H_t(y) = y$, for all $y \in S$, and $H_1(C) \subseteq Y$. Let $\beta = H_1 : C \to Y$ (keep in mind that $\beta(y) = y$, if $y \in S$), and form the reduced mapping cylinder D_β rel S. Then $\dim(D_\beta - Y) \leqslant 2k - n - j + 2$ and $D_\beta \searrow Y$, so that the inclusion $Y \subseteq D_\beta$ is a simple homotopy equivalence. Since $C \subseteq D_\beta$ and $C \searrow T$, the adjunction space $Y_1 = D_\beta \cup_\gamma T$ is simple homotopy equivalent to D_β. (See (5.9) of [17].) Hence, each of the maps $X \to Y \to D_\beta \to Y_1$ is a simple homotopy equivalence. Denote the composition by $h_1 : X \to Y_1$.

Observe that the composition $g' : Y \to D_\beta \to Y_1$ induces the same identifications on Y that g does, so that Y is sent to a subset of Y_1 homeomorphic to $g(Y) \subset M$. Thus, the composition $\gamma_g \circ H : C \times I \to M$, where $\gamma_g : C_g \to M$ is the projection, induces a map $g_1 : Y_1 \to M$ such that $g_1 g' = g$ and $g_1 | g'(Y) (= g(Y))$ is an embedding. Assume g_1 is in general position rel $g(Y)$. Then we have

(a) $g_1 h_1 = f$, and

(b) $\dim S(g_1) \leqslant (2k - n - j + 2) + k - n \leqslant 2k - n - j - 1$,

since $\dim(Y_1 - g(Y)) \leqslant 2k - n - j + 2$, and $n - k \geqslant 3$. The inductive process stops after at most $2k - n$ iterations. □

Using surgery theory, Wall [61] obtains the following embedding theorem, which was proved first for Q simply connected by Casson and Sullivan and by Browder and Haefliger [24]. One can easily see that Theorem 5.8 follows from Theorems 5.9 and 5.10.

THEOREM 5.10 ([61]). *Suppose Q and N are compact PL manifolds of dimensions q and n, respectively, $n - q \geqslant 3$, and $f : (Q, \partial Q) \to (N, \partial N)$ is a homotopy equivalence such that $f | \partial Q$ is a PL embedding. Then f is homotopic,* rel ∂Q, *to a PL embedding.*

Perhaps the first important application of engulfing due to Stallings is his proof of the higher dimensional Poincaré Conjecture [56].

THEOREM 5.11 (Weak Poincaré Conjecture [56]). *Suppose that M is a k-connected, closed PL n-manifold, $n \geqslant 5$, $k = \lfloor n/2 \rfloor$. Then M is topologically homeomorphic to S^n.*

PROOF. Let M_1 be M minus the interior of an n-ball. Then M_1 is also k-connected. Let $K > K_0$ be a triangulation of $M_1 \supset \partial M_1$, let $L = K^{(k)} \cup K_0$, and let $\widetilde{L} < K'$ be the dual $(n-k-1)$-skeleton of K rel K_0. Let N be a regular neighborhood of ∂M_1 in M_1 (a collar), and let B be a small n-ball in $\operatorname{int} M_1$ containing a point $p \notin N$ in its interior. Use 5.2 to get isotopies H^1 and H^2 of M_1, fixed on ∂M_1, such that $H_1^1(N) \supset |L|$ and $H_1^2(B) \supset |\widetilde{L}|$. Using the example of Section 3 and the Regular Neighborhood Theorem, we may assume that $M_1 = H_1^1(N) \cup H_1^2(B)$. The composition $H = H_2^{-1} \circ H_1$ is an isotopy of M_1, fixed on ∂M_1, such that $H(N) \supset M_1 - \operatorname{int} B$. Without loss of generality, $p \notin H(N)$.

Set $M_2 = C\ell(B - H(N))$, and repeat the construction for M_2, obtaining $M_2 = N_2 \cup B_2$, where N_2 is a collar on ∂M_2, $p \notin N_2$, and B_2 is a small n-ball in $\operatorname{int} M_2$ containing p in its interior. After an infinite repetition we obtain $M_1 - \{p\} \cong \partial M_1 \times [0, \infty)$, so that $M - \{p\} \cong \mathbb{R}^n$. Thus M is the one point compactification of $\mathbb{R}^n$, which is topologically homeomorphic to S^n. □

A stronger version of the Poincaré Conjecture, concluding that M is PL homeomorphic to S^n, can be proved for $n \geqslant 6$ using handle theory and the h-cobordism theorem [55,3]. We shall discuss these topics in the next section.

6. Handle theory

Suppose that K is a combinatorial n-manifold with polyhedron M. The combinatorial structure of K provides M with a nice decomposition into PL n-balls, stratified naturally by their "cores", called a *handle decomposition*. Given PL n-manifolds W_1 and W_0, and $J = [-1, 1]$, we say that W_1 *is obtained from* W_0 *by adding a handle of index* p, if $W_1 = W_0 \cup H^{(p)}$, where

$$\left(H^{(p)}, H^{(p)} \cap W_0\right) = \left(H^{(p)}, \partial H^{(p)} \cap \partial W_0\right) \cong \left(J^p \times J^{n-p}, \partial J^p \times J^{n-p}\right).$$

Given a PL homeomorphism $h : (J^p \times J^{n-p}, \partial J^p \times J^{n-p}) \to (H^{(p)}, H^{(p)} \cap \partial W_0)$, we call $h(J^p \times \{0\})$ the *core* of the handle $H^{(p)}$, $h(\partial J^p \times \{0\})$ is the *attaching sphere*, $h(\{0\} \times J^{n-p})$ is the *cocore*, and $h(\{0\} \times \partial J^{n-p})$ is the *belt sphere*. We call h the *characteristic map*, and $f = h|(\partial J^p) \times J^{n-p}$ the *attaching map*.

For example, suppose M is a PL n-manifold without boundary, and K is a combinatorial triangulation of M with first and second derived subdivisions $K' \succ K''$. If σ is a p-simplex of K, then $\operatorname{lk}(\sigma, K) \cong S^{n-p-1}$ so that

$$\begin{aligned} \left(\operatorname{st}(\sigma, K), \sigma\right) &\cong \left(\sigma * \operatorname{lk}(\sigma, K), \sigma\right) \\ &\cong \left(B^p * S^{n-p-1}, B^p\right) \\ &\cong \left(J^p \times J^{n-p}, J^p \times \{0\}\right). \end{aligned}$$

Using a cone construction one in turn sees that

$$\left(\operatorname{st}(\sigma, K), \sigma\right) \cong \left(\operatorname{st}(\hat{\sigma}, K''), \operatorname{st}(\hat{\sigma}, \sigma'')\right),$$

where $\sigma'' = K''|\sigma$.

Denote the PL n-ball, $\operatorname{st}(\hat{\sigma}, K'')$, by B_σ. It is not difficult to see that $B_\sigma \cap B_\tau = \emptyset$ whenever $\dim \sigma = \dim \tau$. Setting $H_p = \bigcup\{B_\sigma\colon \dim \sigma = p\}$, we see that $M = H_0 \cup H_1 \cup \cdots \cup H_n$, where each H_p is a disjoint union of n-balls. (See Figure 3, where B_p denotes a B_σ, $\dim \sigma = p$.) If we set $W_{p-1} = \bigcup_{i<p} H_i$, then for $\dim \sigma = p$, $B_\sigma \cap W_{p-1} = \partial B_\sigma \cap \partial W_{p-1}$ is a regular neighborhood of $\partial \sigma''$ in ∂B_σ; hence, $B_\sigma \cap W_{p-1} \cong S^{p-1} \times J^{n-p}$. Thus, W_p is obtained from W_{p-1} by adding p-handles B_σ, $\dim \sigma = p$, and $M = H_0 \cup H_1 \cup \cdots \cup H_n$.

A *handle decomposition* of a PL n-manifold M is a presentation $M = H_0 \cup H_1 \cup \cdots \cup H_n$, where H_0 is a disjoint union of n-balls, and $W_p = \bigcup_{i \leqslant p} H_i$ is obtained from W_{p-1} by adding p-handles, $1 \leqslant p \leqslant n$. It may be that $W_p = W_{p-1}$, in which case the decomposition has no handles of index p. That is, we allow $H_p = \emptyset$.

If M is a PL n-manifold with boundary ∂M, let C be a regular neighborhood of ∂M in M, $(C, \partial M) \cong (\partial M \times I, \partial M \times \{0\})$. Assume that $C = N(\partial K'', K'')$ for some PL triangulation K of M. Construct the n-balls B_σ as above for $\sigma \notin \partial K$. Then $M = C \cup H_0 \cup \cdots \cup H_n$ as before, and $W_p = C \cup \bigcup_{i \leqslant p} H_i$ is obtained from W_{p-1} by adding p-handles. Notice that any attaching set that meets C meets it in $\dot{N}(\partial K'', K'')$. Any such presentation of M is a *handle decomposition of* M, rel ∂M.

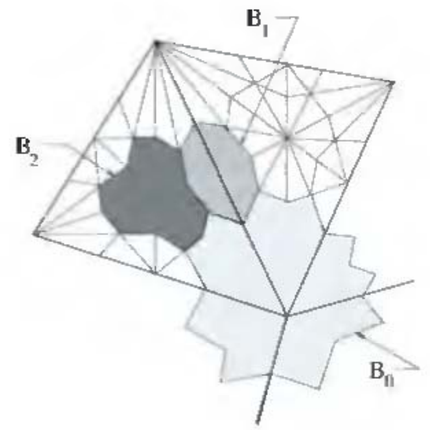


Fig. 3.

Finally, we extend the idea of a handle decomposition to a cobordism between $(n-1)$-manifolds. A *cobordism* is a triple $(W; M, M')$, where W is a PL n-manifold and $\partial W = M \cup M'$, $M \cap M' = \emptyset$. A *handle decomposition* of W, rel M, is a presentation $W = C \cup H_0 \cup H_1 \cup \cdots \cup H_n \cup C'$, where C and C' are regular neighborhoods of M and M' in W, respectively, and $W_p = C \cup \bigcup_{i \leqslant p} H_i$ is obtained from W_{p-1} by adding p-handles.

Dual handle decompositions. Notice that if $W = C \cup H_0 \cup H_1 \cup \cdots \cup H_n \cup C'$ is a handle decomposition of a cobordism $(W; M, M')$, and if $H^{(p)}$ is a p-handle with characteristic map h, then $H^{(p)} \cap (\bigcup_{i>p} H_i \cup C') = h(J^p \times \partial J^{n-p})$. That is, $H^{(p)}$ can be thought of as an $(n-p)$-handle added to $\bigcup_{i>p} H_i \cup C'$. With this point of view, we write $H^{(p)} = \widetilde{H}^{(n-p)}$ and call $\widetilde{H}^{(n-p)}$ the *dual* $(n-p)$*-handle* determined by $H^{(p)}$. Thus, we also get a handle decomposition $W = C' \cup \widetilde{H}_0 \cup \widetilde{H}_1 \cup \cdots \cup \widetilde{H}_n \cup C$, where $\widetilde{H}_p = H_{n-p}$. Dual handle structures are closely related to dual cell structures described in Section 2.

Handle decompositions arising from triangulations of a manifold are generally too large to be of much use, although they often provide a place to get started. The goal is to try to find the simplest possible handle decomposition. For example, if a cobordism $(W; M, M')$ has a handle decomposition with *no* handles, then $W = C \cup C'$ is a product: $(W; M, M') \cong (M \times I; M \times \{0\}, M \times \{1\})$. An obvious necessary condition for this to happen is that the inclusions $M_i \hookrightarrow W$, $i = 0, 1$, are homotopy equivalences. A cobordism $(W; M, M')$ satisfying this condition is called an *h-cobordism* between M and M', or simply an h-cobordism.

h-COBORDISM THEOREM 6.1. *Suppose* $(W; M, M')$ *is a compact h-cobordism,* $\dim W \geqslant 6$, *and* W *is simply connected. Then* $(W; M, M') \cong (M \times I; M \times \{0\}, M \times \{1\})$.

We shall outline a proof of the h-Cobordism Theorem in this section. Our treatment is taken from [50], where many of the omitted details may be found.

Simplifying handle decompositions. For the immediate discussion, we will let $(W; M, M')$ denote a compact cobordism with $\dim W = n$. Our first observation is that "sliding a handle" does not change the topology of the resulting manifold.

LEMMA 6.2. *If $f, g : \partial I^p \times I^{n-p} \to \partial M'$ are ambient isotopic attaching maps, then $W \cup_f H^{(p)} \cong W \cup_g H^{(p)}$ by a homeomorphism that is fixed outside a regular neighborhood (collar) of M'.*

PROOF. By Proposition 3.12, an isotopy of M' extends to W, fixing the complement of a collar on M'. □

LEMMA 6.3. *If $p \leqslant q$, then $(W \cup H^{(q)}) \cup H^{(p)} \cong (W \cup H^{(p)}) \cup H^{(q)}$, with $H^{(p)}$ and $H^{(q)}$ disjoint.*

PROOF. Let S_a be the attaching sphere for $H^{(p)}$ and S_b the belt sphere for $H^{(q)}$. Then $\dim S_a + \dim S_b = (p-1) + (n-q-1) < n-1$ so that S_a can be general positioned to miss S_b in $\partial(W \cup H^{(q)})$. Use 3.14 and Lemma 6.2 to "squeeze" the p-handle so that $H^{(p)} \cap S_b = \emptyset$ as well. Let N be a regular neighborhood of the cocore of $H^{(q)}$ in $W \cup H^{(q)}$ such that $N \cap H^{(p)} = \emptyset$. Since $H^{(q)}$ is also a regular neighborhood, there is an isotopy of $W \cup H^{(q)}$ taking N to $H^{(q)}$. This isotopy slides $H^{(p)}$ off of $H^{(q)}$, so Lemma 6.2 applies to complete the proof. □

As a consequence of Lemma 6.3, we can rearrange the addition of handles to a cobordism W so that the handles are added in nondecreasing order, thereby producing a handle decomposition of W. We now look at circumstances in which handles may be eliminated.

Suppose $W_1 = W \cup H^{(p)} \cup H^{(p+1)}$, $\dim W = n$. Then $H^{(p)}$ and $H^{(p+1)}$ are called *complementary handles* if the attaching sphere S_a of $H^{(p+1)}$ meets the belt sphere S_b of $H^{(p)}$ transversely in a single point. This means that near $x = S_a \cap S_b$, and after an ambient isotopy of the attaching map f for $H^{(p+1)}$, f matches up the product structure on $\partial H^{(p+1)}$ with that on $\partial H^{(p)}$.

LEMMA 6.4. *Suppose that $W_1 = W \cup H^{(p)} \cup H^{(p+1)}$, where $H^{(p)}$ and $H^{(p+1)}$ are complementary handles. Then $W_1 \cong W$ by a PL homeomorphism that is the identity outside a collar on M'.*

PROOF. Let $h : J^p \times J^{n-p} \to H^{(p)}$ be the characteristic map for $H^{(p)}$ and let

$$f : (\partial J^{p+1}) \times J^{n-p-1} \to \partial(W \cup H^{(p)})$$

be the attaching map for $H^{(p+1)}$. Using the Regular Neighborhood Theorem and Lemma 6.2, we may assume that

$$h(J^p \times (\{1\} \times J^{n-p-1})) = f((J^p \times \{1\}) \times J^{n-p-1}),$$

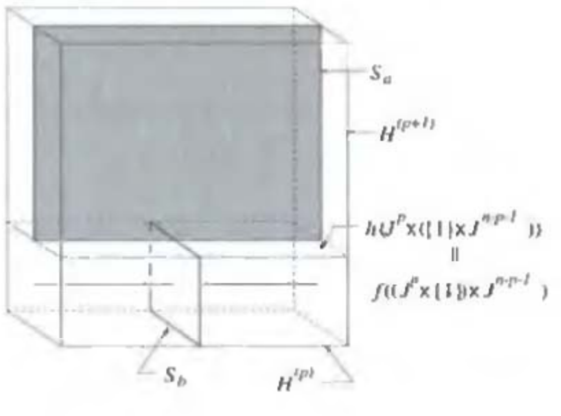


Fig. 4.

and that $f^{-1} \circ h|J^p \times (\{1\} \times J^{n-p-1}) = \text{id}$. (See Figure 4.) Thus, we see that W_1 is obtained from W by attaching an n-ball $B = H^{(p)} \cup H^{(p+1)}$ to M' along an $(n-1)$-ball in ∂B. That is, W_1 shells to W.

The reverse of this process allows one to introduce a cancelling pair of handles to a cobordism.

In general, if $W_1 = W \cup H^{(p)} \cup H^{(p+1)}$, then we can define the *incidence number* $\varepsilon(H^{(p+1)}, H^{(p)})$ as follows. There is a strong deformation retraction of $W \cup H^{(p)}$ onto $W \cup h(I^p \times \{0\})$, where h is the characteristic map for $H^{(p)}$. The composition $W \cup H^{(p)} \to W \cup h(I^p \times \{0\}) \to (W \cup h(I^p \times \{0\}))/W \cong S^p$ gives a mapping $g : W \cup H^{(p)} \to S^p$. If S_a is the attaching sphere for $H^{(p+1)}$, then the restriction $g|S_a$ gives a mapping of $S^p \cong S_a \to S^p$. Choose orientations for (each) S^p, and define $\varepsilon(H^{(p+1)}, H^{(p)})$ to be the degree of this map. Thus, $\varepsilon(H^{(p+1)}, H^{(p)})$ is an integer, which is well-defined up to sign. If S_a and the belt sphere S_b for $H^{(p)}$ are in general position in $\partial(W \cup H^{(p)})$, then they intersect transversely in a finite number of points. If $H^{(p)}$ and $H^{(p+1)}$ are given orientations, then $\varepsilon(H^{(p+1)}, H^{(p)})$ is the algebraic intersection number of S_a and S_b in $\partial(W \cup H^{(p)})$. If $H^{(p)}$ and $H^{(p+1)}$ are a complementary pair of handles, then, clearly, $\varepsilon(H^{(p+1)}, H^{(p)}) = \pm 1$. The next lemma gives conditions under which, up to an ambient isotopy of attaching maps, the converse is true. The proof uses another form of the Whitney Trick and may be found in [50, Chapter 6]. □

LEMMA 6.5 (Handle Cancellation Lemma). *Suppose* $W_1 = (W \cup H^{(p)}) \cup_f H^{(p+1)}$, $2 \leqslant p \leqslant n-4$, $n \geqslant 6$, *M' is simply connected, and* $\varepsilon(H^{(p+1)}, H^{(p)}) = \pm 1$. *Then the attaching map f for $H^{(p+1)}$ is ambient isotopic to an attaching map g such that, in* $W_2 = (W \cup H^{(p)}) \cup_g H^{(p+1)}$, *$H^{(p)}$ and $H^{(p+1)}$ are complementary handles. Thus,* $W_1 \cong W_2 \cong W$.

LEMMA 6.6 (Handle Addition Lemma). *Suppose* $W_1 = W \cup_{f_1} H_1^{(p)} \cup_{f_2} H_2^{(p)}$, *where* $H_1^{(p)} \cap H_2^{(p)} = \emptyset$, $2 \leqslant p \leqslant n-2$, *and M' is simply connected. Then f_1 is ambient isotopic to f_3, where* $[f_3] = [f_1] + [f_2]$ *in* $\pi_p(M')$.

PROOF. If $2 \leqslant p \leqslant n-2$ and M' is connected, we can connect PL embedded $(p-1)$-spheres S_1 and S_2 with a PL "ribbon" $D = g(I \times I^{p-1})$ in M', where g is a PL embedding, $g(I \times I^{p-1}) \cap S_1 = g(\{0\} \times I^{p-1})$ and $g(I \times I^{p-1}) \cap S_2 = g(\{1\} \times I^{p-1})$. (See [50, Chapter 5].) In this way we can add the homotopy classes of $[f_1]$ and $[f_2]$ in $\pi_{p-1}(M')$ (this requires $2 \leqslant p \leqslant n-2$), and if M' is simply connected, the resulting class is independent of D. Inside the boundary of the p-handle $H_1 = H_1^{(p)}$ in W_1, there is a "parallel" copy of its core: $B = h(J^p \times \{x\})$, where $x \in \partial J^{n-p}$ and h is a characteristic map for H_1. Connect the boundary sphere S of B to the attaching sphere S_2 for the handle H_2 with a ribbon D in M'.

Now we have the collapses $B \cup D \cup S_2 \searrow S_2$ and $B \cup D \cup S_2 \searrow S_3 = (S_1 \cup D \cup S_2) - g(I \times \operatorname{int} I^{p-1})$ in $M_2 = \partial(W \cup H_1) - M$. Thus, in M_2, the Regular Neighborhood Theorem provides an ambient isotopy of the attaching map f_1 to $f_3 : S^{p-1} \to M'$ with $[f_3] = [f_1] + [f_2]$ in $\pi_p(M')$. The new handle can be moved off of $H_1^{(p)} \cup H_2^{(p)}$. □

LEMMA 6.7. *Suppose that $(W; M, M')$ is a connected cobordism. Then W has a handle decomposition with no 0- or n-handles.*

PROOF. If a 1-handle $H^{(1)}$ joins C to a 0-handle $H^{(0)}$, then $(H^{(0)}, H^{(1)})$ is a complementary pair. Proceed by induction: if there are any 0-handles, then there is one that is joined to C by a 1-handle. The n-handles in a handle decomposition of W are the 0-handles in the dual decomposition. □

If M is simply connected, one can use a handle decomposition $(W; M, M') = C \cup H_1 \cup \cdots \cup H_{(n-1)} \cup C'$, where $H_p = H_1^{(p)} \cup \cdots \cup H_{n_p}^{(p)}$ is the (disjoint) union of handles of index p, to compute the homology of the pair $(W; M)$. Form the chain complex whose pth chain group, C_p, is generated by the (oriented) p-handles. If $H_i^{(p)}$ is a p-handle, define $\partial(H_i^{(p)}) = \sum_j \varepsilon(H_i^{(p)}, H_j^{(p-1)}) H_j^{(p-1)}$. Observe that if $X \supset Y$ is a polyhedron with $\dim(X - Y) \leqslant p$, and if $f : (X, Y) \to (W, M)$ is a mapping, then f is homotopic, rel Y, to a mapping $g : X \to C \cup H_1 \cup \cdots \cup H_p$. Just proceed inductively: use general position to get $f(X)$ disjoint from the cocores of higher-dimensional handles and then use the handle structure to get $f(X)$ miss the handles themselves. A similar general position argument can also be used to prove the following useful fact.

LEMMA 6.8. *Suppose that $W = C \cup H_0 \cup \cdots \cup H_n \cup C'$ is a handle decomposition, $W^{(p)} = C \cup H_1 \cup H_1 \cup \cdots \cup H_p$, and $M^{(p)} = \partial W^{(p)} - M$. Then*

(a) *$\pi_i(W, W^{(p)}) = 0$ for $i \leqslant p$, and*

(b) *$\pi_i(W, M^{(p)}) = 0$ for $i \leqslant \min\{p, n-p-1\}$.*

LEMMA 6.9. *Suppose that W is connected, $n \geqslant 6$, M is simply connected, and $(W; M, M')$ has a handle decomposition with no handles of index $< p$, $1 \leqslant p \leqslant n-4$. If $H_p(W, M) = 0$, then there is another handle decomposition with no handles of index $\leqslant p$ and with the same number of handles of index $> p + 1$.*

PROOF. Let $W = C \cup H_1 \cup \cdots \cup H_n \cup C'$ be a handle decomposition for W.

Case 1: $p = 1$. Let $H^{(1)}$ be a 1-handle in H_1; $H^{(1)}$ is attached to $M_0 = \partial C - M$. Let α be an arc in $\partial H^{(1)}$ parallel to its core, and let β be a PL arc in C joining the endpoints of α.

Use general position to get α to miss all 2-handles and β to miss the cores of the 1- and 2-handles so that $\gamma = \alpha \cup \beta \subseteq \partial(C \cup H^{(1)}) - M$. By Lemma 6.8 and general position, γ bounds a PL disk D in $M_1 = \partial(C \cup H_1 \cup H_2) - M$. By 4.6 D is locally flat in M_1. Introduce a complementary pair of 2- and 3-handles $H^{(2)}$ and $H^{(3)}$ so that $H^{(2)}$ has attaching sphere ∂D and D lies in the attaching sphere of $H^{(3)}$. Then $H^{(1)}$ and $H^{(2)}$ are complementary handles and may be eliminated. After rearranging handles, we get a new decomposition with $H^{(1)}$ eliminated and a new $H^{(3)}$ introduced.

Case 2: $1 < p \leqslant n - 4$. As $H_p(W, M) = 0$ and there are no handles of index $< p$, the boundary map $\partial : \mathcal{C}_{p+1} \to \mathcal{C}_p$ must be onto. If $H^{(p)}$ is a p-handle, then

$$\sum_i n_i \varepsilon\left(H_i^{(p+1)}, H^{(p)}\right) = 1$$

for some chain $\sum_i n_i H_i^{(p+1)}$ in $\mathcal{C}_{p+1}$. Thus we can use handle addition to get $\varepsilon(H_i^{(p+1)}, H^{(p)}) = 1$ for some $(p+1)$-handle $H_i^{(p+1)}$ and $\varepsilon(H_j^{(p+1)}, H^{(p)}) = 0$ for $j \neq i$. We may then cancel $H^{(p)}$ and $H_i^{(p+1)}$. □

PROOF OF THE h-COBORDISM THEOREM. Let $(W; M, M')$ be a compact h-cobordism, $\dim W \geqslant 6$, with W is simply connected. Let $(W; M, M') = C \cup H_1 \cup \cdots \cup H_{n-1} \cup C'$ be a handle decomposition with no 0- or n-handles. Use Lemma 6.9 to eliminate handles of index $\leqslant n - 4$, leaving only handles of index $(n-3)$, $(n-2)$, and $(n-1)$. Thus the dual handle decomposition only has handles of index 1, 2, and 3. Eliminate the dual 1-handles from the dual decomposition, leaving only dual 2- and 3-handles. Since we are working over the integers, the matrix of the boundary map from $\mathcal{C}_3$ to $\mathcal{C}_2$, which is invertible over $\mathbb{Z}$, may be diagonalized, and the elementary operations may be realized by handle additions. Hence, we arrive at a handle decomposition with only complementary handles in dimensions 2 and 3, which may be cancelled as well, leaving a decomposition with no handles. Hence, W is a product $M \times I$. □

COROLLARY 6.10 (Strong Poincaré Conjecture [3,55]). *If M is a k-connected, closed PL n-manifold, $n \geqslant 6$, $k = \lfloor n/2 \rfloor$, then M is PL homeomorphic to S^n.*

PROOF. Remove the interiors of disjoint, PL n-balls from M. The result is an h-cobordism between boundary spheres, which is a product $S^{n-1} \times I$. □

If W is not simply connected, one must define incidence numbers as elements of the group ring $\mathbb{Z}[\pi_1(M)]$, and the proof of the h-cobordism theorem may break down at the last step, when we eliminate the dual 2- and 3-handles. In this case the matrix of the boundary map is a non-singular matrix over $\mathbb{Z}[\pi_1(M)]$, and it can be diagonalized if, and only if, we are given the additional hypothesis that the inclusion $M \subseteq W$ is a simple homotopy equivalence. An h-cobordism $(W; M, M')$ with this property is called an *s-cobordism.*

THEOREM 6.11 (s-Cobordism Theorem). *Suppose that $(W; M, M')$ is an h-cobordism, $n \geqslant 6$. Then $(W; M, M') \cong (M \times I; M \times \{0\}, M \times \{1\})$ if, and only if, $(W; M, M')$ is also an s-cobordism.*

There are also relative versions of the h- and s-cobordism theorems. Their proofs proceed very much the same as in the absolute case. A *relative cobordism* between $(n-1)$-manifolds M and M' with boundaries is a triple $(W; M, M')$, where W is an n-manifold such that $\partial W = M \cup V \cup M'$, where $M \cap M' = \emptyset$, $\partial V = \partial M \cup \partial M'$, $V \cap M = \partial M$, and $V \cap M' = \partial M'$. A relative cobordism is an h-cobordism (s-cobordism) if all of the inclusions $M, M' \subseteq W$, $\partial M, \partial M' \subseteq V$ are homotopy equivalences (simple homotopy equivalences).

THEOREM 6.12 (Relative s-Cobordism Theorem). *Suppose W is a compact relative h-cobordism, $n \geqslant 6$, such that $V \cong \partial M \times I$. Then $(W; M, M') \cong (M \times I; M \times \{0\}, M \times \{1\})$ if, and only if, $(W; M, M')$ is also an s-cobordism.*

7. Isotopies, unknotting

In this section we explore further the question: Given polyhedra X and Y, when are homotopic embeddings $f, g : X \to Y$ ambient isotopic? The first result is Zeeman's unknotting theorem [65].

THEOREM 7.1 ([65]). *A PL sphere pair (S^n, S^q) or a proper PL ball pair (B^n, B^q), $n - q \geqslant 3$ is unknotted.*

PROOF ([50]). The proof is by induction (eventually) on n. The case $n = 3$ is trivial. The case $n = 4$ follows from Theorem 4.5(i), for sphere pairs, and Theorem 4.7, for ball pairs. Theorem 4.5(ii) gives the result for sphere pairs when $n = 5$, and the Regular Neighborhood Theorem for Pairs can, in turn, be used to show that ball pairs unknot.

Assume, then, that $n \geqslant 6$. Let (B^n, B^q) be a ball pair, $n - q \geqslant 3$. By induction, (B^n, B^q) is locally flat. Hence if N is a regular neighborhood of B^q in B^n, then, by 3.22, (N, B^q) is an unknotted ball pair. Let $W = C\ell(B^n - N)$, $M = N \cap W$. Then W is a relative cobordism between $(M, \partial M)$ and $(M', \partial M')$, where $M' = C\ell((\partial B^n \cap W) - N')$ and N' is a small collar of ∂M in $\partial B^n \cap W$. Since B^q is contractible and $B^n - B^q$ is simply connected, by general position, W is an h-cobordism. Thus $W \cong M \times I$ and so (B^n, B^q) is unknotted.

Suppose (S^n, S^q) is a sphere pair, $n \geqslant 6$, and $n - q \geqslant 3$. Let $K > L$ be a triangulation of $S^n \supset S^q$, and let v be a vertex of L. Then $(\mathrm{st}(v, K), \mathrm{st}(v, L))$ is an unknotted ball pair, as is the complementary pair. Thus (S^n, S^q) is obtained by gluing two unknotted ball pairs together. □

COROLLARY 7.2. *Any proper (n, q)-manifold pair is locally flat, if $n - q \geqslant 3$.*

Since the Whitehead group of $\mathbb{Z}$ is trivial, any h-cobordism between manifolds with fundamental group $\mathbb{Z}$ is an s-cobordism. Thus the proof of Theorem 7.1 works for locally flat sphere pairs in codimension 2, provided the pair is "homotopically unknotted".

THEOREM 7.3. *Suppose (S^n, S^{n-2}), $n \geqslant 6$, is a locally flat sphere pair. Then (S^n, S^{n-2}) is unknotted if, and only if, $S^n - S^{n-2}$ has the homotopy type of a circle.*

We consider next the weaker question: When are isotopic embeddings of X into M ambient isotopic? (See Section 3 for definitions.) The answer is yes whenever the isotopy is "locally extendable". Given an isotopy $F: X \times I \to M \times I$ (or a proper isotopy of pairs) and a point $(x, t) \in X \times I$, F is *locally extendable* at (x, t) if there are neighborhoods V of x in X, U of $F(x, t)$ in M, and J of t in I and a level preserving embedding $h: U \times J \to M \times J$ such that $h(y, t) = (y, t)$ for all $y \in U$ and $h(F(x, s), s) = F(x, s)$ for $(x, s) \in V \times J$. In other words, local collars of $F(X \times \{t\})$ in $F(X \times [t, 1])$ (if $t < 1$) and in $F(X \times [0, t])$ (if $t > 0$) are extendable (locally) to collars of $M \times \{t\}$ in $M \times [t, 1]$ and $M \times [0, t]$. By Theorem 3.19, a locally extendable isotopy is extendable at $X \times \{t\}$ for all $t \in I$, meaning we can choose $V = X$ and $U = M$. This fact together with a standard compactness argument proves the following extension theorem.

THEOREM 7.4. *If X is a compact polyhedron and $F: X \to Y$ is a locally extendable isotopy, then F is ambient.*

If Q and M are PL q- and n-manifolds, respectively, with $n - q \geqslant 3$, then, by Zeeman's Unknotting Theorem, any isotopy of Q in M is locally extendable. This fact, together with Proposition 3.12 establishes the following corollary.

COROLLARY 7.5 ([32]). *Suppose Q and M are PL q- and n-manifolds, respectively, with $n - q \geqslant 3$, and Q is compact. Then any proper isotopy of $(Q, \partial Q)$ in $(M, \partial M)$ is ambient.*

If X and Y are polyhedra and X is compact, a proper PL embedding $f: ((v * X), X) \to (w * Y, Y)$ is *unknotted* if there is a PL homeomorphism $h: w * Y \to w * Y$ fixing Y such that hf is the cone on $f|X$. Lickorish's Cone Unknotting Theorem [38] is a polyhedral analogue of Theorem 7.1. We state it next without proof.

THEOREM 7.6 ([38]). *Suppose that X is a compact $(q - 1)$-dimensional polyhedron and $f: (v * X, X) \to (B^n, S^{n-1})$ is a proper embedding, $n - q \geqslant 3$. Then f is unknotted.*

Theorem 7.6 allows one to prove that a PL isotopy of a polyhedron in a PL manifold in codimension $\geqslant 3$ is locally extendable. Thus, the line of argument above can be used to prove the following isotopy extension theorem due to Hudson [30].

COROLLARY 7.7 ([30]). *Suppose that X is a compact q-dimensional polyhedron, M is a PL n-manifold, $n - q \geqslant 3$. Then any isotopy of X in M is ambient.*

An embedding $F: X \times I \to Y \times I$ satisfying $F^{-1}(Y \times \{i\}) = X \times \{i\}, i = 0, 1$, is a *concordance* of X in Y (from F_0 to F_1). Thus, an isotopy of X in Y is a level-preserving concordance. Hudson [31] obtains the following improvement of Corollary 7.7. Rourke [47] gives a "handle straightening" argument for this result as well.

THEOREM 7.8 ([31]). *Suppose that X is a compact q-dimensional polyhedron, M is a PL n-manifold, $n - q \geqslant 3$. Then concordant embeddings of X in M are ambient isotopic.*

COROLLARY 7.9. *Suppose that Q is a compact PL q-manifold, M is a PL n-manifold, $n-q \geqslant 3$, and $f:(Q,\partial Q) \to (M,\partial M)$ is a (proper) PL embedding such that f is $(2q-n+1)$-connected. If $g:(Q,\partial Q) \to (M,\partial M)$ is a PL embedding that is homotopic to f rel ∂Q, then f and g are ambient isotopic.*

8. Approximations, controlled isotopies

Many of the results of Sections 5, 6, and 7 have "controlled" analogues. In this section we state without proof a few of the basic theorems of this type. The first result is Miller's Approximation Theorem [42].

THEOREM 8.1 ([42]). *Suppose that Q is a PL q-manifold, M is a PL n-manifold, $n-q \geqslant 3$, and $f:(Q,\partial Q) \to (M,\partial M)$ is a topological embedding. Then for every $\varepsilon: Q \to (0,\infty)$, there is a PL embedding $g:(Q,\partial Q) \to (M,\partial M)$ such that $d(f,g) < \varepsilon$.*

(This result was initially announced by Homma [27], but a problem was discovered in his proof by H. Berkowitz. A proof along the lines originally presented by Homma can be found in [9].) Using Miller's theorem for q-cells, Bryant [8] was able to extend Miller's theorem to embeddings of polyhedra.

THEOREM 8.2 ([8]). *Suppose that X is a q-dimensional polyhedron and $f: X \to M$ is a topological embedding into a PL n-manifold. Then for every $\varepsilon: Q \to (0,\infty)$, there is a PL embedding $g: X \to M$ such that $d(f,g) < \varepsilon$.*

The next theorem is an ε version of the isotopy theorems of Section 7. This result is an amalgamation of results due primarily to Connelly [18] and Miller [41], with contributions and improvements due to Cobb [14], Akin [1], and Bryant and Seebeck [10].

THEOREM 8.3 ([18,41]). *Suppose that (X,Y) is a polyhedral pair, $\dim Y < \dim X = q$, M is a PL n-manifold, and $f:(X,Y) \to (M,\partial M)$ is a proper topological embedding, $n-q \geqslant 3$. Then for every $\varepsilon: X \to (0,\infty)$ there is a $\delta: X \to (0,\infty)$ such that if $g_i:(X,Y) \to (M,\partial M)$ are PL embeddings, $i=0,1$, within δ of f, then there is an ε-push H of M such that $Hg_0 = g_1$.*

There are useful variations on Theorem 8.3. For example, if $\dim(X-Y) \leqslant n-3$, $f|Y$ is a PL embedding, and $g_i|Y = f|Y$, $i=0,1$, then one can get an ε-push H of M rel ∂M with $Hg_0 = g_1$.

There are a number of controlled versions of the engulfing theorem (Theorem 5.2), although mostly they have been replaced by Quinn's End Theorem [46]. Bing's article "Radial Engulfing" [5], contains a variety of such theorems the reader is encouraged to survey. We state one such result from [10]. It requires a definition: A subset Y of a space X is *1-LCC* (1-locally co-connected) in X if for each $x \in Y$ and each neighborhood U of x in X, there is a neighborhood V of x in X such that the inclusion $\pi_1(V-Y) \to \pi_1(U-Y)$ is trivial. An embedding $f: Y \to X$ is 1-LCC if $f(Y)$ is 1-LCC in X.

THEOREM 8.4 ([10]). *Suppose $f : X \to M$ is a 1-LCC embedding of a q-dimensional polyhedron X into a PL n-manifold M, $n - q \geqslant 3$, $n \geqslant 5$. Then for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $g : X \to M$ is a PL embedding within δ of f and U is any neighborhood of $g(X)$, then there is an ε-push H of M such that $H_1(f(X)) \subseteq U$.*

If X is not compact then ε and δ are functions of X. The 1-LCC condition allows one to push $f(X)$ off of the 2-skeleton of a neighborhood of $f(X)$ and then close to the dual $(n-3)$-skeleton. One then engulfs the dual $(n-3)$-skeleton with U. These are ε versions of Stallings arguments in [57].

Using Theorems 8.2, 8.3, and 8.4, and an infinite process due to Homma [26] and Gluck [23], one can deduce the following "taming" theorem of Bryant and Seebeck [10].

THEOREM 8.5 ([10]). *Suppose $f : X \to M$ is a 1-LCC embedding of a q-dimensional polyhedron X into a PL n-manifold M, $n - q \geqslant 3$, $n \geqslant 5$. Then for every $\varepsilon > 0$, there is an ε-push H of M such that $H_1 f$ is PL.*

There is a 4-dimensional analogue of this result due to R.D. Edwards (unpublished), obtained from Casson–Freedman handle theory for 4-manifolds [21].

9. Triangulations of manifolds

We conclude this chapter with a discussion of the two most central issues in PL topology: existence and uniqueness of triangulations of topological manifolds. Classically, these questions dealt with PL triangulations of manifolds, although there are obvious related questions concerning triangulations in general. Existence and uniqueness of PL triangulations of manifolds of dimensions $\leqslant 3$ have been known for some time, the case $n = 3$ being due to Moise [44] and Bing [4]. The uniqueness question (the Hauptvermutung) can be asked for polyhedra in general. Milnor was the first to construct examples of polyhedra that are homeomorphic, but not piecewise linearly homeomorphic [43]. Edwards was able to exhibit a non-combinatorial triangulation of S^n, $n \geqslant 5$, by showing that the k-fold suspension of a non-simply connected 3-manifold constructed by Mazur [40] is topologically homeomorphic to S^{k+3}. (See [19].) Edwards and Cannon [11] solved the "Double Suspension Problem" in general by proving the following theorem.

THEOREM 9.1 ([11]). *Suppose that H^n is a PL n-manifold with the homology of S^n. Then, for each $k \geqslant 1$, the polyhedron $S^k * H$ is topologically homeomorphic to S^{k+n+1}.*

Whenever H is not simply connected, which can happen when $n \geqslant 3$, the polyhedron $S^k * H$ is not PL homeomorphic to the standard S^{k+n+1}. Thus, uniqueness of triangulations of topological manifolds fails if one does not require triangulations to be PL. The problem of uniqueness of PL triangulations is more subtle yet.

Suppose that M and N are PL n-manifolds and $h : M \to N$ is a topological homeomorphism. Sullivan's idea [58,59] was to prove that M and N are PL homeomorphic by taking a handle decomposition of M and, inductively, "straightening" their images under h. This idea presents a *handle problem*, that is, a topological homeomorphism $h: B^k \times \mathbb{R}^{n-k} \to V^n$ onto a PL manifold V^n that is PL on a neighborhood of $S^{k-1} \times \mathbb{R}^{n-k}$. The handle can be *straightened* if there is an isotopy H of V^n, fixed on a neighborhood of $S^{k-1} \times \mathbb{R}^{n-k}$ and outside a compact set, such that $H_1 h$ is PL on $h: B^k \times B^{n-k}$. Sullivan showed that, for $n \geqslant 5$, there was a possible $\mathbb{Z}/2$ obstruction to straightening 3-handles [58]. In his solution to the annulus conjecture, Kirby [36] showed how to straighten 0-handles when $n \geqslant 5$. Kirby and Siebenmann (see [37]) proved that k-handles can be straightened provided $n \geqslant 5$ and $k \neq 3$. Whether or not 3-handles could be straightened depended upon the Hauptvermutung for the n-torus $T^n = S^1 \times \cdots \times S^1$. The following result of Hsiang and Shaneson [28], Wall [60] and Casson classifies the PL structures on T^n, showing, in particular, that they are not all equivalent.

THEOREM 9.2. *For $n \geqslant 5$, the set of PL equivalence classes of PL manifolds topologically homeomorphic to T^n, is in one-to-one correspondence with the set of orbits of $(\Lambda^{n-3}\mathbb{Z}^n) \otimes \mathbb{Z}/2$ under the natural action of $GL(n, \mathbb{Z})$. The standard torus corresponds to the zero element under this action.*

($\Lambda^k \mathbb{Z}^n$ denotes the kth exterior algebra on $\mathbb{Z}^n$.) In particular there are non-standard T^5's. The classification implies that if M is a non-standard, or fake, torus, then even covers of M will be standard, while odd covers are not. Kirby [36] used the first fact in his "torus trick" to prove the annulus conjecture, while the second fact is used to disprove the Hauptvermutung. (See [37].)

THEOREM 9.3 ([37]). *Given a PL n-manifold M, $n \geqslant 6$ or $n \geqslant 5$ if $\partial M = \emptyset$, the isotopy classes of PL structures on M are in one-to-one correspondence with the elements of $H^3(M; \mathbb{Z}/2)$.*

The coefficient group $\mathbb{Z}/2$ appears as the homotopy group $\pi_3(\mathrm{TOP/PL})$, where TOP/PL is the fiber of the forgetful map $B_{\mathrm{TOP}} \to B_{\mathrm{PL}}$ of classifying spaces for topological and piecewise linear bundles. Further handle analysis leads to the following existence theorem of Kirby–Siebenmann.

THEOREM 9.4 ([37]). *Given a topological n-manifold M, $n \geqslant 6$, or $n \geqslant 5$ if ∂M has a PL triangulation, there is a well-defined obstruction in $H^4(M; \mathbb{Z}/2)$ to triangulating M as a PL manifold extending the triangulation on ∂M.*

There are, in fact, topological 4-manifolds that do not admit PL triangulations. One such may be constructed using results of Kervaire and Freedman. Kervaire shows [35] that there is a homology 3-sphere H, the Poincaré 3-sphere, that bounds a parallelizable PL 4-manifold M with signature 8. Freedman shows [21] that H bounds a contractible topological 4-manifold M'. $V = M \cup_H M'$ is then a closed 4-manifold with signature 8. A result of Rohlin (see [34]) states that V cannot have a PL triangulation, for, if it did, its signature would be divisible by 16. (A manifold is *parallelizable* if its tangent bundle is a product bundle.)

It is clear that if M has a PL triangulation, then so does $M \times \mathbb{R}^k$ for $k \geqslant 1$. Kirby–Siebenmann prove a very strong converse to this fact for higher dimensional topological manifolds.

THEOREM 9.5 (Product Structure Theorem [37]). *Suppose that M is a topological n-manifold and that $M \times \mathbb{R}^k$, $k \geqslant 1$, is triangulated as a PL manifold. If $n \geqslant 6$, or $n \geqslant 5$ and $\partial M = \emptyset$, then there is a PL triangulation of M inducing an equivalent triangulation on $M \times \mathbb{R}^k$.*

There are relative versions of the Product Structure Theorem, which the reader may find in [37, I, Section 5]. The following, rather obvious corollary has proved useful in applications.

COROLLARY 9.6. *Suppose that M is a topological n-manifold with boundary, $n \geqslant 6$, such that* $\operatorname{int} M$ *has a PL triangulation. Then M has a PL triangulation.*

In particular, any topological, n-dimensional submanifold (with boundary) of a PL n-manifold M, $n \geqslant 6$, has a PL structure. Likewise, if a compact subset C of a topological n-manifold, $n \geqslant 6$, has vanishing Čech cohomology in dimension 4, then naturality of the obstruction in Theorem 9.4 implies that sufficiently small manifold neighborhoods of C have PL structures.

The question as to whether a topological n-manifold has a triangulation (PL or not) has been investigated extensively by Galewski and Stern. (See, e.g., [22].) Let θ_3^H denote the group obtained from oriented PL homology 3-spheres, under the operation of connected sum, #, modulo those that bound acyclic PL 4-manifolds. There is a homomorphism $\mu : \theta_3^H \to \mathbb{Z}/2$ (the Kervaire–Milnor–Rohlin map) defined by $\mu(H) = [\sigma(W)/8]$, where $\sigma(W)$ denotes the signature of any parallelizable 4-manifold W with boundary H. If H is the Poincaré homology 3-sphere, then $\mu(H) = 1$, so that μ is surjective.

THEOREM 9.7 ([22]). *Suppose M is a topological n-manifold, $n \geqslant 6$ or $n \geqslant 5$ if ∂M is triangulated. Then there is an element $t_M \in H^5(M, \partial M; \ker \mu)$ such that $t_M = 0$ iff there is a triangulation of M compatible with the given triangulation on ∂M. Moreover, the set of concordance classes of triangulations of M* rel ∂M *is in one-to-one correspondence with the elements of $H^4(M, \partial M; \ker \mu)$.*

Triangulations K_0 and K_1 of M are *concordant* if there is a triangulation K of $M \times I$ restricting to K_i on $M \times \{i\}$, $i = 0, 1$.

Galewski and Stern [22] and Matumoto [39] have shown that all compact topological n-manifolds ($n \geqslant 6$ or $n \geqslant 5$ if ∂M is triangulated) can be triangulated if there is a homology 3-sphere H such that $\mu(H) = 1$ and $H\#H$ bounds a parallelizable 4-manifold with signature 0. At the time of this writing it is unknown whether such a 3-manifold exists or whether every topological n-manifold, $n \geqslant 5$, can be triangulated. Casson, however, has found a topological 4-manifold that cannot be triangulated [13].

References

[1] E. Akin, *Manifold phenomena in the theory of polyhedra*, Trans. Amer. Math. Soc. **143** (1969), 413–473.

[2] P. Alexandroff and H. Hopf, *Topologie, I*, Springer, Berlin (1945).

[3] D. Barden, *The structure of manifolds*, Doctoral thesis, Cambridge Univ. (1963).

[4] R.H. Bing, *An alternative proof that 3-manifolds can be triangulated*, Ann. of Math. (2) **69** (1959), 37–65.

[5] R.H. Bing, *Radial engulfing*, Conf. on the Topology of Manifolds, J.G. Hocking, ed., Prindle, Weber, & Schmidt, Boston (1968).

[6] R.H. Bing and J.M. Kister, *Taming complexes in hyperplanes*, Duke Math. J. **31** (1964), 491–511.

[7] M. Brown, *Locally flat embeddings of topological manifolds*, Ann. of Math. (2) **75** (1962), 331–341.

[8] J.L. Bryant, *Approximating embeddings of polyhedra in codimension three*, Trans. Amer. Math. Soc. **170** (1972), 85–95.

[9] J.L. Bryant, *Triangulation and general position of PL diagrams*, Topology Appl. **34** (1990), 211–233.

[10] J.L. Bryant and C.L. Seebeck, *Locally nice embeddings in codimension three*, Quart. J. Math. Oxford (2) **21** (1970), 265–272.

[11] J.W. Cannon, *Shrinking cell-like decompositions of manifolds in codimension three*, Ann. of Math. (2) **110** (1979), 83–112.

[12] S. Cappell and J. Shaneson, *Piecewise linear embeddings and their singularities*, Ann. of Math. (2) **103** (1976), 163–228.

[13] A. Casson, unpublished work.

[14] J.I. Cobb, *Taming almost PL embeddings in codimensions* 3, Abstract 68T-241, Notices Amer. Math. Soc. **15** (2) (1968), 371.

[15] M.M. Cohen, *Simplicial structures and transverse cellularity*, Ann. of Math. **85** (1967), 218–245.

[16] M.M. Cohen, *A general theory of relative regular neighborhoods*, Trans. Amer. Math. Soc. **136** (1969), 189–229.

[17] M.M. Cohen, *A Course in Simple Homotopy Theory*, Springer, New York (1970).

[18] R. Connelly, *Unknotting close polyhedra in codimension three*, Topology of Manifolds, J.C. Cantrell, ed., Markham, Chicago (1970).

[19] R.J. Daverman, *Decompositions of Manifolds*, Academic Press, Orlando (1986).

[20] A.I. Flores, *Über n-dimensionale Komplexe die im* $\mathbb{R}^{2n+1}$ *Absolut selbstverschlungen sind*, Ergib. Math. Kolloq. **6** (1935), 4–7.

[21] M.H. Freedman, *The topology of four-dimensional manifolds*, J. Differential Geom. **17** (1982), 357–453.

[22] D. Galewski and R. Stern, *A universal 5-manifold with respect to simplicial triangulations*. Geometric Topology (Proc. Georgia Topology Conf., Athens, GA, 1977), Academic Press, New York (1979), 345–350.

[23] H. Gluck, *Embeddings in the trivial range*, Ann. of Math. (2) **81** (1965), 195–210.

[24] A. Haefliger, *Knotted spheres and related geometric problems*, Proc. I.C.M. (Moscow, 1966), Mir (1968), 437–444.

[25] M.W. Hirsch, *On tubular neighbourhoods of manifolds*, Proc. Cambridge Philos. Soc. **62** (1966), 177–183.

[26] T. Homma, *On the imbedding of polyhedra in manifolds*, Yokohama Math. J. **10** (1962), 5–10.

[27] T. Homma, *A theorem of piecewise linear approximations*, Yokohama Math. J. **14** (1966), 47–54.

[28] W.-C. Hsiang and J. Shaneson, *Fake tori*, Topology of Manifolds, J.C. Cantrell, ed., Markham, Chicago (1970), 18–51.

[29] J.F.P. Hudson, *P. l. embeddings*, Ann. of Math. (2) **85** (1967), 1–31.

[30] J.F.P. Hudson, *Piece Linear Topology*, W.A. Benjamin, New York (1969).

[31] J.F.P. Hudson, *Concordance, isotopy, and diffeotopy*, Ann. of Math. (2) **91** (1970), 425–448.

[32] J.F.P. Hudson and E.C. Zeeman, *On combinatorial isotopy*, Publ. I.H.E.S. (Paris) **19** (1964), 69–94.

[33] M.C. Irwin, *Embeddings of polyhedral manifolds*, Ann. of Math. (2) **82** (1965), 1–14.

[34] M.A. Kervaire, *On the Pontryagin classes of certain SO(n)-bundles over manifolds*, Amer. J. Math. **80** (1958), 632–638.

[35] M.A. Kervaire, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. **144** (1969), 67–72.

[36] R.C. Kirby, *Stable homeomorphisms and the annulus conjecture*, Ann. of Math. (2) **89** (1969), 575–582.

[37] R.C. Kirby and L.C. Siebenmann, *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations*, Princeton Univ. Press, Princeton, NJ (1977).
[38] W.B.R. Lickorish, *The piecewise linear unknotting of cones*, Topology **4** (1965), 67–91.
[39] T. Matumoto, *Variétés simpliciales d'homologie et variétés topologiques métrisables*, Thesis, Univ. de Paris-Sud, Orsay (1976).
[40] B. Mazur, *A note on some contractible* 4*-manifolds*, Ann. of Math. (2) **73** (1961), 221–228.
[41] R.T. Miller, *Close isotopies on piecewise-linear manifolds*, Trans. Amer. Math. Soc. **151** (1970), 597–628.
[42] R.T. Miller, *Approximating codimension* 3 *embeddings*, Ann. of Math. (2) **95** (1972), 406–416.
[43] J. Milnor, *Two complexes that are homeomorphic but combinatorially distinct*, Ann. of Math. (2) **74** (1961), 575–590.
[44] E.E. Moise, *Affine structures in* 3*-manifolds, V: The triangulation theorem and Hauptvermutung*, Ann. of Math. (2) **56** (1952), 96–114.
[45] J.R. Munkres, *Topology, A First Course*, Prentice-Hall, Englewood Cliffs, NJ (1975).
[46] F. Quinn, *Ends of Maps, I*, Ann. of Math. (2) **110** (1979), 275–331.
[47] C.P. Rourke, *Embedded handle theory, concordance and isotopy*, Topology of Manifolds, J.C. Cantrell, ed., Markham, Chicago (1970).
[48] C.P. Rourke and B.J. Sanderson, *An embedding without a normal microbundle*, Invent. Math. **3** (1967), 293–299.
[49] C.P. Rourke and B.J. Sanderson, *Block bundles: I*, Ann. of Math. (2) **87** (1968), 1–28.
[50] C.P. Rourke and B.J. Sanderson, *Introduction to Piecewise-Linear Topology*, Springer, Berlin (1972).
[51] A. Scott, *Infinite regular neighborhoods*, J. London Math. Soc. **42** (1963), 245–253.
[52] A. Shapiro, *Obstructions to the imbedding of a complex in euclidean space*, Ann. of Math. (2) **66** (1955), 256–269.
[53] L.C. Siebenmann, *Infinite simple homotopy types*, Indag. Math. **32** (5) (1970), 479–495.
[54] I.M. Singer and J.A. Thorpe, *Lecture Notes on Elementary Topology and Geometry*, Scott, Foresman, and Co., Glennview, IL (1967).
[55] S. Smale, *Genralized Poincaré's conjecture in dimension* > 4, Ann. of Math. (2) **74** (1961), 391–466.
[56] J. Stallings, *The piecewise linear structure of euclidean space*, Proc. Cambridge Philos. Soc. **58** (1962), 481–488.
[57] J. Stallings, *The embedding of homotopy types into manifolds*, Preprint.
[58] D. Sullivan, *Triangulating homotopy equivalences*, Doctoral dissertation, Princeton Univ. (1966).
[59] D. Sullivan, *On the Hauptvermutung for manifolds*, Bull. Amer. Math. Soc. **73** (1967), 598–600.
[60] C.T.C. Wall, *Homotopy tori and the annulus theorem*, Bull. London Math. Soc. **1** (1969), 95–97
[61] C.T.C. Wall, *Surgery on Compact Manifolds*, Academic Press, London (1970).
[62] C. Weber, *L'élimination des points doubles dans le cas combinatoire*, Comment. Math. Helv. **41** (1966–67), 179–182.
[63] J.H.C. Whitehead, *Simplicial spaces, nuclei, and m-groups*, Proc. London Math. Soc. **45** (1939), 243–327.
[64] H. Whitney, *The self-intersection of a smooth n-manifold in* $2n$*-space*, Ann. of Math. (2) **45** (1944), 220–246.
[65] E.C. Zeeman, *Unknotting spheres*, Ann. of Math. (2) **72** (1960), 350–361.
[66] E.C. Zeeman, *Seminar on Combinatorial Topology* (notes), I.H.E.S. (Paris) and Univ. of Warwick (Coventry) (1963–1966).

CHAPTER 6

Geometric Group Theory*

J.W. Cannon

Department of Mathematics, Brigham Young University, Provo, UT 84602-6539, USA
E-mail: cannon@math.byu.edu

Contents

*This research was supported in part by an NSF research grant.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

0. Introduction

Almost twenty years ago, W.P. Thurston established a firm link between classical geometry and low-dimensional topology in dimension 3. The basic reference is [47].

These notes are currently available from the Geometry Center at the University of Minnesota in Minneapolis. The first four chapters of the notes have been carefully reworked and expanded and are available in book form [48].

Thurston's work made it very clear that the dominant invariant in the study of 3-manifolds is the fundamental group and that the nature of the fundamental group is very much a function of the geometry associated with both the group and the manifold through his work. His results suggested that geometry might serve as a guiding principle in the study of infinite group theory.

Such thoughts were, of course, not entirely new, for combinatorial group theory had largely begun because of topological and geometric work of Max Dehn early in the century. The fundamental reference is [20].

However, combinatorial group theory had largely abandoned, or at least somewhat obscured, its topological and geometric roots in an attempt to distill the combinatorial essence of the arguments. Thurston's work brought renewed emphasis to those roots.

The hopes raised by Thurston's theory, that concentration on geometry might enrich infinite group theory once more, have been amply borne out during the two decades intervening now between Thurston's original work and the present. There has been a veritable explosion in the application of geometric methods to infinite group theory.

This chapter is intended as an entry point into the resulting theory.

Section 1 is motivational and outlines some of the paradigmatic work in geometric group theory.

Section 2 explains, or gives specific reference to, the foundational concepts connecting infinite group theory and geometry. In order to feel at home in geometric group theory, one must come to understand at least the following concepts:

- What is a group presentation? What are generators and relators?
- How does a presentation define a group?
- How does one obtain a presentation for the fundamental group of a space?
- What is a geometry?
- What is a geometric action of a group on a geometry?
- What is the Cayley graph of a group?
- How does one construct a Cayley graph by Todd–Coxeter coset enumeration?
- In particular, what is a free product with amalgamation, what is an HNN extension, and how does one construct their Cayley graphs?
- If a group acts geometrically on a geometry, how uniquely is the geometry determined? (Answer: the geometry is unique up to quasi-isometry.)

Section 3 is dedicated to the proposition propounded by Marshall Hall: "Theorems change, but examples remain." In Section 3.1 we give a family of examples of finitely presented groups naturally associated with geometries. In Section 3.2 we then describe in very abbreviated fashion the prototypical, standard geometries in dimensions $\leqslant 3$ with which the sample groups are associated. Then finally in Section 3.3 we explain how to construct the Cayley graphs of most of the groups in Section 3.1. We would hope that the reader would at first ignore Section 3.3 and attempt to construct the Cayley graphs with the help only of the material in Section 2. To do so would be a difficult, lengthy, perhaps impossible exercise; but, if carried out with even partial success, it would firmly place the reader in the subject. After one has the appropriate Cayley graphs, one can use them as discrete models for the geometries which they approximate. Much can be learned about the geometries by thinking about the associated Cayley graphs. In fact, the author's own entry into the field came from his own self-imposed task of constructing "all" of the Cayley graphs of the groups described in the wonderful book [18].

Though we did not succeed in constructing all of these Cayley graphs, we did observe fascinating things about groups which have served as the basis for much of our work.

Section 4 outlines three of the most important abstract classes of infinite groups studied by means of geometric group theory, the word-hyperbolic or negatively curved groups, the automatic groups, and the CAT(0) groups.

Section 5 enumerates some of the geometries one might study as motivation for further work in geometric group theory.

And finally, Section 6 directs the reader toward the existing work on the numerous topological and geometric invariants that have application to the study of infinite groups.

1. Paradigmatic work in geometric group theory

Every scientific subject is driven by its best and most suggestive results. In this section we describe five of the many beautiful results of geometric group theory which have been extremely influential in setting the direction of this subject.

1.1. *Dehn's work on the fundamental group of a topological surface*

Max Dehn [20] introduced the three fundamental problems of combinatorial (and geometric) group theory.

They are of course, the *word problem*, which Dehn called the *identity problem*, the *conjugacy problem*, which Dehn called the *transformation problem*, and the *isomorphism problem*. Here are Dehn's definitions.

THE WORD PROBLEM. *Suppose an element of the group is given as a product of generators of the group. One should give a method in order to decide in a finite number of steps whether this element represents the identity element of the group or not.*

THE CONJUGACY PROBLEM. *Suppose two elements S and T of the group are given. A method is sought to decide the question whether S and T are conjugate in the group; that is, does there exist an element U in the group such that $S = UTU^{-1}$?*

THE ISOMORPHISM PROBLEM. *Two groups are given. One should decide whether they are isomorphic or not.*

Dehn gave beautiful solutions to all three problems when the groups involved are the fundamental groups of compact 2-dimensional surfaces. We shall only describe Dehn's solution of the word problem, and only that in the case where the surface is closed, orientable, of genus 2. The fundamental group has the following presentation.

$$G = \langle x_1, y_1, x_2, y_2 \mid x_1 y_1 x_1^{-1} y_1^{-1} x_2 y_2 x_2^{-1} y_2^{-1} = 1 \rangle.$$

We call the long relator, which is the product $[x_1, y_1][x_2, y_2]$ of two commutators, R. We consider all cyclic permutations of R and its inverse as well as all trivial relators of the form $z \cdot z^{-1}$, where z can be any of the generators or their inverses. We call the resulting finite collection of words the set S of *fundamental relators* for G. Dehn's algorithm for solving the word problem is the following. Let w be any word in the generators of G or their inverses. Look in w for an occurrence of a word M such that one of the elements of S has the form Mm, where M is a word that is longer than the (possibly empty) word m. Replace the occurrence of M by the word m^{-1}. Continue until no more such shortening replacements can be performed. Then w represents the identity element of G if and only if the resulting maximally shortened word is the empty word.

Dehn's solution to the conjugacy problem used his solution to the word problem and had a similar flavor.

Dehn's solution to both problems used his familiarity with non-Euclidean geometry and the fact that the universal cover of his surface was in fact the non-Euclidean plane.

Dehn's solutions were generalized in two directions. First of all, the combinatorial underpinning of the argument was distilled; the consequence was *small cancellation theory*. A fundamental reference is [32, Chapter V].

In the late 1970's we discovered that Dehn's algorithms could be generalized almost verbatim to all groups acting nicely on non-Euclidean geometry, and in fact to any groups that acted sufficiently like non-Euclidean hyperbolic geometry. The reference is [11].

Gromov gave an appropriate definition of the class of groups acting sufficiently like non-Euclidean geometry in the fundamental paper [29].

Gromov showed that small cancellation theory could be made very geometric and returned the subject to the geometric grounds where it began.

1.2. *Gromov's theorem on groups of polynomial growth*

Consider the free Abelian group $Z \oplus Z$ on two generators and the free non-Abelian group $Z * Z$ on two generators. If one expresses each element as a word of minimal possible length in terms of the standard generators and their inverses, then the free Abelian group has exactly $4n$ elements of minimal length n for $n \geqslant 1$, while the free non-Abelian group has exactly $4 \cdot 3^{n-1}$ elements of minimal length n for $n \geqslant 1$. We say that the free Abelian group has *polynomial growth* since there is a polynomial in n which bounds the number of elements of length n and that the free non-Abelian group has *exponential growth* because there is no such polynomial and in fact the number of elements is bounded below by an exponential function of n. Gromov has proved the following theorem [28]:

THEOREM. *A finitely presented group G has polynomial growth if and only if G has a nilpotent subgroup of finite index.*

In addition to its intrinsic beauty, the result and its proof are of interest to topologists and geometers for at least three reasons.

(1) The result shows that the simplest of geometric properties, namely the geometric growth rate of a group, can have profound influence on the algebraic structure of a group.
(2) Gromov describes how a sequence of metric spaces can converge to another metric space.
(3) Gromov uses, at a critical juncture, the solution to Hilbert's fifth problem about the topological structure of Lie groups.

1.3. *Stallings's theorem on the number of ends of a group*

Let X denote a locally compact, connected metric space. If K is a compact subset of X, then we say that a component C of $X \setminus K$ is *unbounded* if the closure of C in X is noncompact. We define $e(X, K)$ to be the number of unbounded components. If K' is a compact set which is larger than K, then each unbounded component of $X \setminus K$ will contain at least one unbounded component of $X \setminus K'$ so that $e(X, K') \geqslant e(X, K)$. We define the *number of ends* of X to be the supremum over all K of the numbers $e(X, K)$. The reader will easily verify that the real line has 2 ends, the plane and all Euclidean spaces of higher dimension have only one end.

One can assign a number of ends to a finitely generated group in two steps. First one assigns to the group and any of its finite generating sets a locally compact, connected metric space, namely its Cayley graph (see Section 2.3). The Cayley graph has a number of ends by the previous paragraph, and an easy argument shows that the number of ends does not depend on the generating set. We call this the number of ends of the group. For the free-Abelian group of rank 2 with its usual two-element generating set, the Cayley graph is the square unit lattice in the Euclidean plane; it clearly has 1 end. For the free non-Abelian group of rank 2 with its usual two-element generating set, the Cayley graph is an infinite tree with four edges meeting at every vertex; it clearly has infinitely many ends. It has long been known that a group has either 0, 1, 2, or infinitely many ends. The groups with 0 ends are finite. The generic case seems to be the groups with 1 end. Stallings [45] characterizes groups with 2 or infinitely many ends as follows.

THEOREM. *A finitely generated group G has exactly two ends if and only if G has an infinite cyclic subgroup of finite index. A finitely generated group G has infinitely many ends if and only if G can be factored in one of the two following ways*:

(1) *G is a free product with finite amalgamating subgroup where this amalgamating group is properly contained in both factors and of index > 2 in at least one factor*;
(2) *G is an HNN extension amalgamated over a finite subgroup which is properly embedded in the base group.*

(*Free products with amalgamation and HNN extensions, and their Cayley graphs, are described at the end of Section* 2.)

Stallings remarks that he discovered the proof of this theorem by contemplating Papakyriakopoulos's proof of the sphere theorem.

1.4. *Mostow's rigidity theorem*

It is sometimes possible to endow a topological manifold with a Riemannian metric of constant curvature. For example, a torus of any dimension can be endowed with a flat Euclidean metric. A 2-dimensional surface of genus g at least two can be endowed with a Riemannian metric of constant curvature -1; as a consequence, the universal cover of such a surface is naturally isometric with non-Euclidean hyperbolic geometry and the covering translations within this geometry are hyperbolic isometries. In dimension 2, the dimension we have just been discussing, there is a large space of such metric structures. This space is called Teichmüller space. It is homeomorphic with Euclidean space of dimension $6g-6$. This space has been studied extensively in classical complex variables and conformal mapping. In higher dimensions however Mostow noted that such metrics of constant negative sectional curvature occur much less frequently and then are much more rigidly determined. Here is a weak version of Mostow's rigidity theorem [38].

MOSTOW RIGIDITY THEOREM. *Suppose that M_1 and M_2 are closed manifolds of dimension $n \geqslant 3$ endowed with Riemannian metrics of constant sectional curvature equal to -1. If M_1 and M_2 have isomorphic fundamental groups, then M_1 and M_2 are isometric. Thus all of the differential geometric invariants of M_i (volume, spectrum of geodesic lengths, conformal structure at infinity, etc.) are in fact topological invariants of M_i.*

The proof makes strong use of the space at infinity, quasiconformality, ergodicity, weak metric equivalences called *quasi-isometry*, all of these notions being strongly geometric. The theorem suggests that in dimensions at least 3 algebra may in appropriate settings determine topology, and even Riemannian geometry, completely.

1.5. *Thurston's geometrization conjecture*

The geometric flexibility of topological surfaces in dimension 2 suggests that geometry will not play a major role in the study of manifolds. Mostow's rigidity theorem is a first corrective in showing that, in dimensions higher than 2, algebra, topology, and geometry can be very tightly associated. Thus Thurston could say approximately the following in his address to the International Congress of Mathematicians in Helsinki in 1978: "A compact manifold is nearly determined by its fundamental group. We can almost identify the *geometric* study of manifolds with the *algebraic* study of finitely presented groups."

Thurston analyzed the geometric structure of Haken 3-manifolds (his results have since been generalized to a larger class of 3-manifolds). He identified all of the geometries (simply connected, complete Riemannian manifolds, homogeneous under isometry) which could serve as universal covering space for a closed 3-manifold with Riemannian metric in such a way that the covering transformations were by isometry. There were exactly eight. (We shall describe them in a later section, Section 3.2.) He then showed that every Haken 3-manifold had a canonical decomposition into pieces each of which admitted a complete Riemannian metric of finite volume with universal cover isometric to one of the eight geometries. We summarize this result in the following abbreviated fashion:

THURSTON'S GEOMETRIZATION THEOREM. *Every Haken* 3*-manifold admits an essentially unique geometric structure.*

One good reference to the Thurston work is [37].

The Thurston work is explicitly about topology, but it, together with Mostow's rigidity theorem, suggests firm connections with group theory and outlines implicitly a path to be taken in understanding the groups which arise in 3-dimensional topology, namely to analyze the individual groups associated with the 8 geometries and how those groups can be assembled to obtain all 3-manifold groups.

2. The foundations of geometric group theory

In this section we address the fundamental issues that must be understood if one is to be comfortable with geometric group theory. These issues were outlined point by point in the introduction.

- What is a group presentation? What are generators and relators?

A group presentation has the form $G = \langle C \mid R \rangle$. In this presentation, C denotes a set of symbols or letters called the *generating set* for the presentation. Each letter $c \in C$ is paired with another letter c^{-1} called its *inverse* letter. Sometimes we assume for convenience that $C = C^{-1}$; at other times we simply describe C by giving only one element of each element-inverse pair. In either case, a *word* in the alphabet C is a finite sequence of elements of $C \cup C^{-1}$. The set R is a collection of words in the alphabet C and is called the *defining relator set* for the presentation. The symbol G refers to the group defined by the presentation as described in the next section. Sometimes it is convenient to define a relator set by equations of the form $v = w$, where v and w are words; this equation, which is called a *relation* rather than a relator, is to be associated with the *relator* $v^{-1} \cdot w$. That is, relations are associated in a natural way with relators.

- How does a presentation define a group?

If $f : C \cup C^{-1} \to H$ is a function into a group H such that, for each $c \in C \cup C^{-1}$, we have $f(c^{-1}) = f(c)^{-1}$, then there is a natural extension of f which takes the collection of all words in the alphabet f into the group H: one simply takes a word $c_1 \cdot c_2 \cdots c_n$ to the element $f(c_1) \cdot f(c_2) \cdots f(c_n)$. We also denote the extended map by f.

We consider the category whose objects are functions $f : C \cup C^{-1} \to H$, as in the previous paragraph, but with the additional property that the extended function f sends each word in the defining relator set R to the identity element of H. We define a morphism in this category from object $f_1 : C \cup C^{-1} \to H_1$ to object $f_2 : C \cup C^{-1} \to H_2$ to be a group homomorphism $h : H_1 \to H_2$ such that, for each $c \in C \cup C^{-1}$, $h(f_1(c)) = f_2(c)$.

The *group defined by the presentation* $G = \langle C \mid R \rangle$ is any initial object in this category; we recall that an *initial object* of a category is an object which admits a unique morphism into any other object of the category. Initial objects are clearly unique up to isomorphism.

Here is a precise construction of the group defined by the presentation $G = \langle C \mid R \rangle$. Let W denote the set of all words in the generating set C. Let T denote the set of *trivial words*, that is words of the form $c \cdot c^{-1}$, where $c \in C \cup C^{-1}$. Declare two words w and w' to be equivalent if w' can be obtained from w by either inserting or deleting at some point a copy of a word from $R \cup T$. Extend this notion of equivalence to an equivalence relation $\sim$. Let $G = W/\sim$ denote the set of equivalence classes. Define the product of two equivalence classes $[w]$, represented by the word $w \in W$, and $[x]$, represented by the word $x \in W$, to be the class $[wx]$ represented by the concatenation wx of w and x. It is easy to check that multiplication is well-defined, defines a group structure on G, and that the function $f : C \cup C^{-1} \to G$ which takes c to $[c]$ is an initial object in the category defined in the preceding paragraphs. A word equivalent to the empty word is called a consequence of R.

- How does one obtain a presentation for the fundamental group of a space?

We assume that the reader is quite familiar with the fundamental group. A good basic reference is [35].

It is easy to obtain a presentation for the fundamental group of a connected space if it happens to be the underlying space of a simplicial complex. Indeed, assume that X is such a space, triangulated, with base point x_0 a vertex of the triangulation. Collapse a maximal subtree of the 1-skeleton to the base point x_0. Then the image of the 1-skeleton of X is a bouquet of loops, each loop supplying a generator for the fundamental group. In the process, each 2-cell is attached along its boundary to the bouquet by an attaching map which can be realized as a word in the generators and their inverses. These words supply the defining relators of a presentation. If the complex is finite, then the presentation is finite.

If a space has reasonably simple local structure, then it is likewise easy to obtain a presentation for the fundamental group. We start with a space X that is connected and locally path connected. We say that an open cover is *2-set simple* if each element is connected and each loop which lies in the union of any 2 sets of the cover is contractible in the space. We say that the space X is *2-set simple* if it has an open cover which is 2-set simple. One could also define the notion of *1-set simple* space; such a space is called *semi-locally simply connected*. Semi-local simple connectivity is important as the condition necessary and sufficient for the existence of a universal covering space. We shall see in the next theorem that it is easy to find a presentation for the fundamental group of a space that is 2-set simple. The result will make use of the notion of *nerve of a covering*. If U is a covering, then the nerve $N(U)$ has one vertex $v(u)$ for each element $u \in U$. A set $\{v(u_0), v(u_1), \ldots, v(u_k)\}$ of $k + 1$ vertices spans a k-simplex of the nerve if and only if the corresponding elements $u_0, \ldots, u_k$ have nonempty intersection.

THEOREM. *Suppose X is a space that is connected and locally path connected, with base point x_0, suppose X is 2-set simple, and suppose U is an open cover of X which is 2-set* *simple. Then the fundamental group of the nerve $N(U)$ of the open covering U is isomorphic with the fundamental group of X.*

COROLLARY. *If M is a compact metric topological manifold with boundary, then the fundamental group is finitely presented.*

PROOF. It is clear that M has a finite open covering which is 2-set simple. The nerve is a finite simplicial complex, therefore has finitely presented fundamental group. □

PROOF OF THE THEOREM. Pick for $N(U)$ a base vertex $v(u_0)$ where $x_0 \in u_0 \in U$.

We establish the following facts in turn. (1) Loops in the nerve $N(U)$ can be reasonably easily mapped to loops in X. (2) The image of a loop is well-defined up to homotopy and so has well-defined image in the fundamental group of X. (3) A nullhomotopic loop maps to a nullhomotopic loop; consequently the map on loops is well-defined on homotopy classes of loops in $N(U)$ and defines a homomorphism from fundamental group of $N(U)$ to the fundamental group of X. (4) The homomorphism is surjective and has trivial kernel, hence is an isomorphism.

(1) Let P denote the set of closed edge paths in $N(U)$ based at $v(u_0)$. For each loop $p \in P$ we construct a loop q in X based at x_0 as follows. Represent p by a sequence $(v(u_0), v(u_1), \ldots, v(u_n), v(u_0))$ of vertices defining the endpoints of successive edges of p. For each i, pick $y_i \in u_i$, with $y_0 = x_0$. Since $v(u_i)$ and $v(u_{i+1})$ span an edge of $N(U)$, the sets u_i and u_{i+1} intersect, so we may take a path q_i in their union which joins y_i to y_{i+1}. We then define q to be the path which is the concatenation $q_0 q_1 \cdots q_n$ of the paths q_i.

(2) We claim that the homotopy class of the path q is well-defined in the sense that it does not depend on the choices of points $y_i \in u_i$ and paths q_i. Suppose that y_i' and q_i' are other admissible choices. Choose paths r_i in u_i joining y_i to y_i'. Then the paths $q_i' r_{i+1}^{-1} q_i^{-1} r_i$ lie in $u_i \cup u_{i+1}$, hence are nullhomotopic in X since the cover U is 2-set simple. These nullhomotopies, taken together, build a homotopy from $q_0 q_1 \cdots q_n$ to $q_0' q_1' \cdots q_n'$, so that the homotopy class of q is well-defined.

(3) Suppose that p is nullhomotopic. Then we see that q is also nullhomotopic as follows. Let $g : B^2 \to N(U)$ be a singular disk bounded by the nullhomotopic edge path p. We may assume that the domain disk B^2 is triangulated and that g is simplicial. We may view q as a map from $S^1 = \partial B^2$ into X with S^1 divided into segments exactly as in the map g and with segments corresponding to edges of p mapping to paths q_i. Our task is to extend this map to all of B^2 as a map $h : B^2 \to X$. We extend the map to the entire 1-skeleton of B^2 in the same way that we defined q in (1). It remains therefore only to define the map h on the 2-simplexes of B^2. We therefore consider a 2-simplex $v_1 v_2 v_3$ of B^2. Since g is simplicial, if u_i, with $i = 1, 2, 3$, satisfies $g(v_i) = v(u_i)$, then the three sets u_i intersect. We map the barycenter b of the 2-simplex to a point in the intersection. We then map the segment bv_i to a path in u_i which joins $h(b)$ and $h(v_i)$. The three segments bv_i divide the 2-simplex into three 2-simplexes, with the boundary of each mapped into the union of two elements of U. Since the cover U is 2-set simple, each of these three boundary maps is contractible. Hence we may extend h to each of the three 2-simplexes. This completes the definition of h.

We conclude that the homotopy class of q depends only on the homotopy class of p so that we have an induced mapping $f:\pi_1(N(u), u_0) \to \pi_1(X, x_0)$. We also conclude that this mapping is a homomorphism.

(4) Suppose that $h: S^1 \to X$ is a loop based at the base point x_0. Partition the loop h into subpaths $q_0, q_1, \ldots, q_n$, each q_i having image in a single element u_i of U, with $u_0 = u_n$ being the base element of U. Then $p = (v(u_0), v(u_1), \ldots, v(u_n))$ is a loop whose image is homotopic to h. Therefore, the homomorphism f is onto.

Suppose now that a loop p is given, with image loop q. Suppose that the path q is nullhomotopic as demonstrated by a map $h: B^2 \to X$ with boundary q. We wish to show that p is also nullhomotopic. It will follow that the homomorphism f is 1 to 1.

It is, of course, permissible to change p by a homotopy and thereby to permit a slight simplification of q. Our aim is a normal form for p such that each edge is either degenerate in $N(U)$ with (possibly nondegenerate) image q_i in X which maps into a single element of U or is nondegenerate in $N(U)$ with image q_i that is degenerate (constant) in X. We can attain that goal by the following simple stratagem. Replace the edge $v(u_i)v(u_{i+1})$ in p by three edges, $v(u_i)v(u_i)$, $v(u_i)v(u_{i+1})$, and $v(u_{i+1})v(u_{i+1})$, the first and the last of these three being degenerate edges. However, in defining q, map the first degenerate edge to a path in u_i from y_i to a point z_i in the intersection of u_i and u_{i+1}, map the nondegenerate edge to the constant path at z_i, and map the final degenerate edge to a path in u_{i+1} from z_i to y_{i+1}. We may therefore, in particular, assume that each edge of q is contained in a single element of U.

Since each edge of q is contained in a single element of U, we may triangulate B^2 without changing the boundary triangulation in such a way that each simplex maps into a single element of U.

We now define the nullhomotopy $g: B^2 \to N(U)$ of p simplex by simplex.

An interior vertex z of B^2 has image $h(z) \in X$ which lies in some element u of U. Define $h(z) = v(u)$, where $v(u)$ is the vertex of $N(U)$ corresponding to u.

Let $e = v_1v_2$ denote an interior edge of B^2 and let b denote its barycenter. The image $h(e)$ lies in a single element u of U. Define $g(b) = v(u)$. Since u contains both $h(v_1)$ and $h(v_2)$, u intersects the elements of U corresponding to both $g(v_1)$ and $g(v_2)$. Hence both $g(b)g(v_1)$ and $g(b)g(v_2)$ are edges of $N(U)$ so that we may extend g to take the half-edges bv_1 and bv_2 onto these edges of $N(U)$.

Let $t = v_1v_2v_3$ denote one of the triangles of B^2 and let b denote its barycenter. The image $h(t)$ lies in a single element u of U. Define $g(b) = v(u)$. Since u contains all three of $h(v_1)$, $h(v_2)$, and $h(v_3)$, it follows that u intersects the elements of U previously chosen with respect to all the faces of t. Consequently, the entire boundary of t is already mapped by g into the star of $g(b)$ in $N(U)$. We may thus extend g to all of t by a cone map.

This completes the construction of $g: B^2 \to N(U)$ which shrinks the closed path p and completes the proof of the theorem. □

- What is a geometry?

For our purposes, a *geometry* is a topological space endowed with a *proper path metric*; a *path metric* is a metric such that the distance between each pair of points is realized as the length of some path in the space joining those points; the metric is *proper* if closed metric balls of finite radius are compact.

What are the prototypical geometries? There is one standard geometry in dimension 1, namely the real line R with its Euclidean metric $d(x, y) = |x - y|$. There are three standard geometries in dimension 2, and eight in dimension 3; we shall enumerate them in Section 3.2. We consider them prototypical because

(1) They are simply connected, connected, metrically homogeneous Riemannian manifolds.

(2) They admit a geometric action (defined immediately hereafter) by some group.

- What is a geometric action of a group on a geometry?

An *action* $G \times S \to S$ of a group G on a set S is a function $(g, s) \mapsto gs$ such that, for each $g_1, g_2 \in G$ and for each $s \in S$, $g_1(g_2 s) = (g_1 g_2)s$ and $\mathrm{id} \cdot s = s$. An action is *isometric* if S is a metric space and, for each $g \in G$ and for each $x, y \in S$, $d(gx, gy) = d(x, y)$. An action is *cocompact* if the orbit space S/G is compact. An action is *properly discontinuous* if, for each compact set $K \subset S$, the set $\{g \in G \mid K \cap gK \neq \emptyset\}$ is finite. An action is *geometric* if S is a geometry and the action is isometric, cocompact, and properly discontinuous. One should view the defining properties of a geometric action in the following way. The isometry condition connects the group action with the metric. The compactness condition ensures that the group is relatively large with respect to the size of the geometry. The proper discontinuity property ensures that the group is not too large with respect to the size of the geometry. That is, the group is compatible with the geometry and is neither too small nor too large, but rather just right. We shall see in Section 2.4 that all geometries on which a group can act geometrically are equivalent under the equivalence relation of *quasi-isometry*.

Suppose now that X is a compact geometry, not simply connected, but semi-locally simply connected (1-set simple in our previous terminology, consequently also 2-set simple since the space is metric, with finitely presented fundamental group). Then we may lift the local metric on X to a path metric on the universal covering space $\widetilde{X}$ of X. Then $\widetilde{X}$ becomes a geometry. The fundamental group acts on this geometry. The action is isometric precisely because the metric on $\widetilde{X}$ is lifted from a metric on X. The action is cocompact because the compact space X is the quotient of $\widetilde{X}$ under the action. The action is properly discontinuous by a fundamental property of covering transformations. That is, the finitely presented group $\pi_1(X)$ (base point suppressed) acts geometrically on $\widetilde{X}$.

Properties of the group $\pi_1(X)$ are reflected in properties of the geometry $\widetilde{X}$. Here is a sample well-known theorem.

THEOREM. *A group G acts geometrically on an n-connected geometry, $n \geqslant 0$, if and only if there is an Eilenberg–MacLane space $K(G, 1)$ with finite $(n+1)$-skeleton.*

The case $n = 0$ implies that a group must be finitely generated in order to act geometrically on a geometry. The case $n = 1$ implies that a group must be finitely presented in order to act geometrically on a simply connected geometry. And so forth.

One proof of the theorem appears in the appendix of the paper [13].

- What is the Cayley graph of a group?

2.1. *Cayley graphs or Dehn Gruppenbilder*

Let G denote a group and $C = C^{-1}$ a generating set for G. Then the *Cayley graph* or *Dehn Gruppenbild* $\Gamma = \Gamma(G, C)$ for G with respect to the generating set C is a 1-dimensional CW complex defined as follows. The vertices of Γ are the elements of G. If v is a vertex and if $c \in C$, then we associate with v and c a labelled and directed edge denoted $(v, c, v \cdot c)$ which has v as its *initial endpoint*, $v \cdot c$ as its *terminal endpoint*, and c as its *color* or *label*. The directed edge $(v, c, v \cdot c)$ is identified topologically with a second edge $(v \cdot c, c^{-1}, v)$ in an order reversing fashion. The second edge is considered the *inverse* or *reverse* of the first. The two directed edges define the same single undirected topological edge joining v and $v \cdot c$. The resulting graph can be turned into a path metric space by assigning a length of 1 to each edge.

There is a natural action of G on its Cayley graph Γ. Here is the action on vertices v and edges $(v, c, v \cdot c)$, respectively:

$$(g, v) \mapsto gv;$$
$$\big(g, (v, c, v \cdot c)\big) \mapsto (gv, c, gv \cdot c).$$

By assigning a linear structure to one edge in each orbit of edges and transferring that structure to all edges in the orbit by the action, we obtain a path metric that is invariant under the action.

THEOREM. *The Cayley graph Γ is a geometry if and only if the generating set C is finite. If the Cayley graph Γ is a geometry, then the natural action $G \times \Gamma \to \Gamma$ is a geometric action.*

Note that this result is compatible with the result of the preceding subsection where it was noted that a necessary and sufficient condition that G act geometrically on some (connected) geometry is that the group G be finitely generated.

Topologists may find the following alternative description of Γ enlightening. Pick a presentation for G: $G = \langle C \mid R \rangle$ where $C = C^{-1}$ is the given generating set for G and the set R is a defining set of relators for G. Choose some point b as base vertex for a CW complex. Attach to b a bouquet B of loops, one for each element-inverse pair c, c^{-1} from C. For each relator $r \in R$ (r being a finite word in the generators C) attach a disk $D(r)$ to B so that the boundary circle of $D(r)$ has image in B which traces out the word r in B. Let K denote the resulting 2-dimensional CW complex. Then the fundamental group of K is G. Let $\widetilde{K}$ denote the universal cover of K. Then $\widetilde{K}$ is called the *Cayley complex* of the presentation, and its 1-skeleton is the Cayley graph $\Gamma = \Gamma(G, C)$.

The alert reader will note that there is a slight incompatibility in our two descriptions of the Cayley graph when one of the generators $c \in C$ is its own inverse. In the graph description one obtains edges of the form $(v, c, v \cdot c)$ with inverse edge $(v \cdot c, c, v)$, the pair forming a single undirected edge. In the topological universal-cover description one obtains two edges spanned by a disk which gives the relator $c^2 = \mathrm{id}$. Collapsing such disks to an arc recovers the graph description from the topological description.

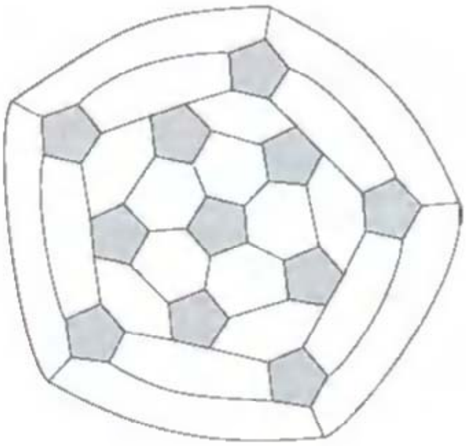


Fig. 1. Soccerballohedron, the Cayley graph of the alternating group on five letters.

As an illustration without supporting proofs we take the group defined by the presentation

$$G = \langle x, y \mid x^2 = y^5 = (xy)^3 = 1 \rangle.$$

Here is the Cayley complex with disks collapsed corresponding to the relator $x^2 = 1$; the topological space underlying the Cayley complex is the 2-sphere S^2 which we view as the 1-point compactification of the Euclidean plane, with the compactifying point at infinity.

The 1-skeleton naturally subdivides the 2-sphere into twelve pentagons (shaded, the outermost one containing the point at infinity) swimming in a sea of twenty hexagons. The pentagons correspond to the relator $y^5 = 1$, the hexagons represent the relator $(xy)^3 = 1$, and the edges connecting different pentagons represent the edges to which the relators labelled $x^2 = 1$ have been collapsed.

The configuration pictured above has been enjoyed for millenia. The sports enthusiast will immediately recognize the pattern as being that of the soccer football, or, as it is affectionately known, the *soccerballohedron*.

The most striking conclusion to be drawn from the figure is that the Cayley graph of a group, realized geometrically, has a natural (in this case, spherical) shape.

What information can be read easily from this geometric representation?

Order: there are 60 vertices so that the group has order 60.

Commutativity: start at a vertex and follow paths labelled xy, then yx; the two paths end at different vertices; hence the group is non-Abelian.

Cosets: the twelve pentagons represent the twelve left cosets in G of the subgroup $\langle y \rangle$.

Multiplication table: a standard multiplication table would require $60^2 = 3600$ entries; all of that information is encoded in the graph and its directed adjacencies.

Symmetry: every vertex sits in the graph exactly in the same manner of very other vertex; it is therefore irrelevant which vertex is considered the identity vertex.

Recognizing the Cayley graph. In giving the sample Cayley graph above, we have obviously ignored a number of critical issues. How does one construct the Cayley graph of a group? When one encounters a graph, how does one recognize that it is in fact a Cayley graph of some group? that it is the Cayley graph of the group in which we are interested? We explain these issues in this subsection.

The most daunting result states that it is, in general, impossible to construct a prescribed finite chunk of an infinite Cayley graph algorithmically; this impossibility result is known as the *unsolvability of the word problem for finitely presented groups*. Two enlightening references concerning this result are [43, Chapter 12, pp. 277–321], [32, Chapter IV, Section 7].

Nevertheless, many Cayley graphs can be constructed algorithmically, and there is a natural algorithm called *Todd–Coxeter coset enumeration* which works for all groups in the sense that it will construct each finite chunk of the Cayley graph correctly in a finite amount of time. The difficulty is that, in general, one does not know when to stop and say that the algorithm has succeeded; the algorithm grinds away and constructs larger and larger chunks of the graph correctly, but it takes so much time that *no computable sequence of stopping times grows large enough rapidly enough to serve as a predictor which says when enough is enough.*

We outline here the theoretical underpinnings which ensure the correctness of the Todd–Coxeter algorithm and which allow one to construct and recognize the Cayley graph of many finitely presented groups. In fact, as we shall see in a later section, the Todd–Coxeter algorithm succeeds nicely for the generic finitely presented group, namely for all negatively curved (that is, word-hyperbolic) groups (Section 4.1). All of the results which we cite in this section are elementary. Readers who will develop proofs for each of them will soon find themselves completely comfortable with Cayley graphs. The proofs appear in an unpublished expository paper of the author [14] which he will be happy to supply to interested parties.

We give a few of the proofs here as hints to the reader who is attempting to work out the proofs alone.

A *colored graph* Γ with *color set* $C = C^{-1}$ consists of a set V of vertices and a set $E \subset V \times C \times V$ of directed and labelled edges (v, c, w) such that, if (v, c, w) is an edge, then its *inverse* or *reverse* (w, c^{-1}, v) is also an edge. This inverse pair of edges is said to define the same topological undirected edge.

The standard example of a colored graph is the *coset graph* (Schreier or Todd–Coxeter diagram): let G denote a group, $C = C^{-1}$ a generating set, and H a subgroup of G; set $\Gamma = (V, C, E)$ where $V = \{Hg \mid g \in G\}$ and $E = \{(Hg, c, Hg \cdot c) \mid g \in G,\ c \in C\}$. If H is the trivial subgroup of G, then Γ is the Cayley graph.

What properties of colored graphs single out the coset graphs and Cayley graphs?

The connectedness property. Since C is a generating set for G, coset and Cayley graphs are clearly connected: there is a path of edges labelled by appropriate generators which take any element of G (or coset of H) to any other element of G (or coset of H). We call a graph which satisfies the connectedness property a *connected* graph.

The permutation property. Since multiplication $(Hg, c) \mapsto Hgc$ by c permutes the cosets of H, there is one and exactly one edge labelled c entering and exiting from each vertex of Γ. We call a graph which satisfies the permutation property a *permutation* graph.

Characterization theorem for coset graphs. A colored graph is a coset graph if and only if it is a connected permutation graph.

What distinguishes Cayley graphs among coset graphs? Suppose that we have a colored graph $\Gamma = (V, C, E)$ that is a connected permutation graph. Given any finite word w in the generating set $C = C^{-1}$ and any vertex $v \in V$, there is a unique edge path in Γ which begins at v and whose successive edge labels are the letters of w. We say that w is a *relator* at v if this unique edge path is a closed path.

The homogeneity property. Since G acts isometrically on its Cayley graph Γ by left multiplication and since there is an element of G which takes any vertex v to any other vertex v' (namely, left multiplication by $v' \cdot v^{-1}$), any relator w at v is also a relator at v'. We summarize this by saying that Γ is *homogeneous*.

Characterization theorem for Cayley graphs. A coset graph is a Cayley graph if and only if it is homogeneous. As defining relators one may take any collection of relators of which all other relators are consequences.

At this point we refer the reader back to the section where we constructed the group defined by given generators and relators. Suppose that we have a Cayley graph Γ. For which group is it a Cayley graph? By the characterization theorem for Cayley graphs, the graph Γ has the same relator set at each of its vertices, namely precisely those words are relators which are consequences of the defining relator set. The Cayley graph $\Gamma = (V, C, E)$ has as its relator set at each vertex precisely the words that are consequences of the relators R. This property of relator sets characterizes the Cayley graph of $G = \langle C \mid R \rangle$.

The observations of the previous paragraph supply the key to the construction of the Cayley graph Γ for G: we must construct a connected permutation graph such that the relators at each vertex are precisely those words that are consequences of the defining relators R of some presentation $G = \langle C \mid R \rangle$. The following lemmas indicate how to control the relators of a constructed graph.

THE BASE POINT LEMMA. *Let Γ denote a connected colored graph and let v and w denote vertices of Γ. Then every relator at v is a consequence of the relators at w.*

PROOF. Let r be a relator at v. Let p be a path from w to v in the (connected) graph Γ. Then prp^{-1} is a relator at w. Then r is equivalent to $p^{-1}prp^{-1}p$ via trivial relators, which is equivalent to $p^{-1}p$ via the relator prp^{-1} at w, which is equivalent to the empty word via trivial relators. □

THE TREE LEMMA. *Let Γ denote a connected colored graph that is separated by each of its edges. Then all of the relators in Γ are consequences of the trivial relators T.*

PROOF. Hint: if r is a relator at vertex v, then the path which begins at v and is labelled by the letters of r has vertices which are furthest from v; take the first edge leading to such a vertex and show that edge is cancelled by the very next edge. □

THE WEDGE LEMMA. *Let Γ_1 and Γ_2 denote two disjoint connected colored graphs and let v_1 and v_2 denote vertices of Γ_1 and Γ_2, respectively. Let Γ denote the graph formed by* *identifying v_1 with v_2. Then every relator in Γ is a consequence of the union of the relators of Γ_1 and Γ_2.*

PROOF. Hint: use the base point lemma to see that one needs only to consider the relators at v_1 and v_2. □

Let Γ denote a connected colored graph and let $\sim$ denote an equivalence relation on the vertices of Γ. We may then take a quotient Γ' of the graph Γ by taking the equivalence classes as vertices and identifying edges that have the same initial vertex, the same terminal vertex, and the same label.

THE QUOTIENT LEMMA. *Let R denote a set of words. Let Γ denote a connected colored graph. Let $\sim$ denote a relation (not necessarily an equivalence relation) such that, if vertices v_1 and v_2 are related, then there is a word $r \in R$ such that either $v_1 \cdot r = v_2$ or $v_2 \cdot r = v_1$; that is, there is a path joining v_1 and v_2 with label r. Extend $\sim$ to an equivalence relation on the vertices of Γ. Then every relator of the quotient graph Γ' is a consequence of the union of the set R and the relators of Γ.*

PROOF. Let $e_1 \cdots e_n$ denote an edge path in the quotient graph that is a relator. Lift each edge e_i to an edge f_i in the original graph Γ. The terminal vertex of f_i is joined to the initial vertex of f_{i+1} (or, of f_n and f_0) by a path g_i which is labelled by a product of elements of $R \cup R^{-1}$. The word $e_1 \cdots e_n$ is therefore equivalent to $e_1 f_1 e_2 f_2 \cdots e_n f_n$ via elements of $R \cup R^{-1}$, and the latter word is a relator in Γ. □

THE RECOGNITION THEOREM. *A coset graph $\Gamma = (V, C, E)$ is the group graph defined by presentation $\langle C \mid R \rangle$ provided*

(1) *Each $r \in R$ is a relator at each vertex of Γ, and*
(2) *Each relator of Γ is a consequence of R.*

PROOF. We indicate the argument. In order to show that Γ is the desired Cayley graph, we must show that the relators at each vertex are *precisely* the consequences of R.

By (2) there are no superfluous relators at any vertex. Hence we can finish the proof by showing that every consequence w of the defining set R of relators is in fact a relator at every vertex v of Γ.

Fix a vertex v of Γ and fix a consequence w of the defining relator set R. Then w can be transformed into the empty word by a finite number of operations which either insert or delete an element of $R \cup T$, where T is the set of trivial words. We may assume by induction (which begins with the empty word) that w is only one transformation away from a word which is a relator at v. Let us suppose, for example, that $w = w_1 \cdot r \cdot w_2$ where $r \in R \cup T$ and where $w_1 \cdot w_2$ is a relator at v. Since Γ is a permutation graph (has the permutation property), every trivial relator is a relator at each vertex of Γ; taking this fact together with property (1) of the hypothesis, we see that r is necessarily a relator at each vertex of Γ, in particular at the terminal vertex $v \cdot w_1$ of the path which begins at v and is labelled by the word w_1. But this means that we can delete from the path labelled w which begins at v the subpath which begins at $v \cdot w_1$ and is labelled r without disrupting the path.

We pass thereby to the path which begins at v and is labelled by $w_1 \cdot w_2$. By inductive hypothesis, this path is a closed path (a relator). It follows that the original path labelled by w is also closed. The complementary case where w is formed by deleting a relator r is handled similarly. In this case too we conclude that w is a relator at v. These observations complete the proof of the recognition theorem. □

- How does one construct a Cayley graph by Todd–Coxeter coset enumeration?

The Todd–Coxeter algorithm. We give only the details of the construction of the Cayley graph by coset enumeration with the subgroup H being the trivial subgroup. We leave the modifications necessary for $H \neq 1$ to the reader.

We assume a group presentation $\langle C \mid R \rangle$ given. Our task is to construct a connected permutation graph which has as its relators at each vertex precisely the consequences of R.

We proceed by induction. The induction begins with a single vertex which we think of as the identity or base vertex v_0 of Γ. The graph Γ will be a limit of graphs Γ_n, each with a base vertex denoted v_0. The limit is taken in the following sense: for each integer $k > 0$ there will be an integer $n(k)$ after which the neighborhoods or balls in Γ_n of radius $\leqslant k$ no longer change; then Γ may be taken as the union of these unchanging balls, individually unchanging but growing with k.

The graph Γ_{n+1} is created from the graph Γ_n by a three-stage process.

Stage 1, the expansion stage. In order that Γ be a permutation graph, we must have at least one edge with each label which emanates from each vertex of Γ. If this property fails at any vertex of Γ_n, we call that vertex *incomplete*. We need to complete each vertex of Γ_n by adding each missing edge. Each individual edge of course is a tree and so has only trivial relators by the tree lemma. Wedging these edges to Γ_n adds only consequences of $R \cup T$ by the wedge lemma. In the process, of course, we create new incomplete vertices, but that problem will be fixed at the next expansion stage. We call the new colored graph $\Gamma_n(1)$.

Stage 2, the relator stage. We can assume inductively that the words of R are relators at most of the vertices of $\Gamma_n(1)$. However, for some of the more recently added vertices this need not be the case. If v is a vertex of $\Gamma_n(1)$ and if there is a labelled path in $\Gamma_n(1)$ which begins at v and which has $r \in R$ as its label, then we may take a quotient graph of $\Gamma_n(1)$ which identifies v with $v \cdot r$. By the quotient lemma, every relator of the quotient graph is a consequence of the relators R. Having carried out this process for every vertex of $\Gamma_n(1)$ we obtain a colored graph $\Gamma_n(2)$. As this process is repeated, we ensure condition (1) of the recognition theorem at more and more of the vertices of our graph.

Stage 3, the permutation stage. The expansion stage, when repeated, ensures that the trivial relators will be relators at every vertex of our graph. This condition is not, however, sufficient to ensure that the result will be a permutation graph because we may find that we have more than one edge with a given label emanating from the same vertex. We need to fold those edges together so that they become one edge. The vertices which we wish to identify thereby are joined by a path which has a trivial relator as its label; therefore taking such a quotient will, by the quotient lemma, introduce no undesired consequences. We therefore identify vertices which are joined by paths labelled by trivial relators. As a consequence of these identifications we make sure that no more than one edge emanates from each vertex with any given label. One should note, however, that the creation of such a quotient may introduce new vertices that have more than one edge emanating from a given vertex with a given label. Thus one may obtain a cascade of identifications. It is in these cascades that the uncertainty in the construction of Γ occurs. The fact that the word problem cannot in general be solved lies precisely in the existence of these unpredictable cascading collapses. At any rate, after the collapsing has stopped, we obtain a colored graph $\Gamma_n(3)$ which we also call Γ_{n+1}.

Note that, for any integer k, there are only finitely many colored graphs of radius $\leqslant k$ which have no more than one edge emanating from any given vertex with a given label. Hence the balls of radius $\leqslant k$ in the graphs Γ_n, which at worst for $n \geqslant k$ may become smaller by the taking of quotients but never become bigger, must eventually stabilize. Thus the limit Γ exists. This graph is easily seen, by the recognition theorem, to be the Cayley graph of G.

Final remarks on the Todd–Coxeter process. The construction of any infinite Cayley graph by the Todd–Coxeter process takes, of course, infinitely many steps. The remarkable fact is that it is often easy to recognize the algorithmic patterns that occur in the process so that one can *see* the entire infinite result in clear detail just as one can *see* the entire unit lattice in the Euclidean plane. The pictures one obtains are almost without exception compellingly beautiful.

Exercises on the construction of Cayley graphs

EXERCISE 1. Construct the Cayley graph of the soccer ball by the Todd–Coxeter algorithm. Be sure to use the appropriate lemmas and the recognition theorem to verify that you have constructed the right graph.

EXERCISE 2. Construct the Cayley graphs of the groups given in Section 3.1.

- In particular, what is a free product with amalgamation, what is an HNN extension, and how does one construct their Cayley graphs?

The most useful constructions of combinatorial group theory are *free products with amalgamation* and *HNN extensions*. It is a fine exercise to use the tools prepared in the previous section to describe the Cayley graphs of these constructs. One can use the recognition theorem to verify that the descriptions are correct. Thereby one can immediately verify the important normal form theorems for elements of these groups (Britten's lemma for the HNN case).

DEFINITION. Let G_1 and G_2 denote groups having isomorphic subgroups H_1 and H_2. Consider presentations

$$G_i = \langle C_i \mid R_i \rangle$$

for these groups having the special property that in the generating set C_i for G_i there is a subset D_i which lies in and generates H_i. Assume further that the sets D_1 and D_2 correspond bijectively according to the given isomorphism between H_1 and H_2: $d_1(\alpha) \in D_1$

corresponding to $d_2(\alpha) \in D_2$, where the symbol α is simply an index from some suppressed index set A. Then the *free product with amalgamation* has presentation

$$G_1 *_{(H_1=H_2)} G_2 = \langle C_1 \cup C_2 \mid R_1 \cup R_2,\ d_1(\alpha) = d_2(\alpha) \rangle.$$

Constructing the Cayley graph of the free product with amalgamation. We must construct a connected permutation graph on the colors $C_1 \cup C_2$ which satisfies the two requirements of the recognition theorem for Cayley graphs. That is, we want the defining relators to be relators at each vertex, but we want every relator in the graph to be a consequence of the defining relators.

We build the Cayley graph recursively.

We start with a copy of the Cayley graph of G_1. The graph is connected and has precisely one edge into and out of each vertex labelled with the colors of C_1. Every relator is, of course, a consequence of the relators R_1 and every relator in the defining set R_1 is satisfied at each vertex.

As defects, however, we do not have edges labelled with the colors of C_2, nor have we satisfied the relators of R_2 nor the new relators of the form $d_1(\alpha) = d_2(\alpha)$.

The principle step of our recursive description deals with a single left coset of H_1 in G_1. We see these cosets by examining the subgraph with edges labelled by elements of D_1. This subgraph falls into components corresponding precisely to the left cosets gH_1 of H_1 in G_1.

By homogeneity of the graph, each of these cosets looks in the graph precisely like each of the others and each vertex looks like each of the others as well. We take one coset gH_1 and one vertex v of that coset. Corresponding to that coset, we take one copy of the Cayley graph of G_2 and one vertex w of that graph. We identify v and w. By the wedge lemma, all relators in the new graph are consequences of the relators in the old graph. We now take the quotient which identifies the coset gH_1 with the coset of H_2 in the graph of G_2 which contains the vertex w. All of the identifications made are consequences of the new relators $d_1(\alpha) = d_2(\alpha)$. One obtains a new graph which is a quotient of the two original Cayley graphs, the two graphs intersecting at the vertices corresponding to exactly one coset in each.

The resulting graph has been improved in the following way. At each element of the two cosets, both generating sets appear, in and out as required by the permutation property, and all of the defining relators are satisfied.

The remaining cosets in the two graphs have not been affected and suffer all of the old defects.

However, the recursion now becomes clear. For each incomplete coset of G_1 repeat the process with a new copy of the graph of G_2. As a consequence, all of the cosets of the original copy of the Cayley graph of G_1 will be complete. However, each new copy of the graph of G_2 will have cosets that are incomplete. Complete each of those with a new copy of the graph of G_1. Now there will be incomplete cosets in the new copies of G_1. Complete those cosets. And so forth.

Take the union of all the finite stages. This graph satisfies all of the properties of the recognition theorem. Notice that the recursion puts the copies of the graphs of G_1 and G_2 together in the pattern of a tree. Following paths through the tree in the order of the construction gives a natural description of the elements of the free product with amalgamation.

DEFINITION. The other construction that is basic is called the HNN extension (after *Higman, Neumann, and Neumann*). It deals with a single group G which has two isomorphic subgroups H_1 and H_2. We take a presentation

$$G = \langle C \mid R \rangle$$

such that C has subsets D_1 generating H_1 and D_2 generating H_2, again in such a manner that the sets D_1 and D_2 correspond under the given isomorphism of H_1 with H_2, $d_1(\alpha)$ corresponding to $d_2(\alpha)$. The HNN extension then has one additional generator t, called a *stable letter*, and a presentation

$$G*_{(H_1=H_2)} = \langle C, t \mid R,\ td_2(\alpha)t^{-1} = d_1(\alpha) \rangle.$$

Constructing the Cayley graph of an HNN extension. Again one starts with a copy of the Cayley graph of G and considers a left coset of H_1 in G. Corresponding to that (incomplete) coset, one takes another copy of the graph of G. In that new copy, one considers a single left coset of the group H_2. One connects the old copy with the new copy via a whole family of edges labelled t which join elements of g_1H_1 with elements of g_2H_2 in a manner corresponding to the isomorphism given between H_1 and H_2. By homogeneity it does not really matter what vertices are considered as the identity vertex; the result will be the same, and all relators will be consequences of the new relators and the old.

One repeats this process with each left coset of H_1. However, in the case of the HNN extension we do not thereby complete the graph at this first copy of G; for we have edges labelled t exiting from all of these vertices, but we do not have any edges labelled t entering these vertices. One must also complete the left cosets of H_2 in the original copy of the graph of G.

Each new copy of the graph of G has incomplete cosets of both H_1 and of H_2 which must be completed.

Again, after infinitely many steps one obtains the Cayley graph of the HNN extension. Again one notes that it has been put together recursively in essentially a treelike fashion. Tracing paths through the defining treelike structure gives Britten's lemma which supplies normal forms for the elements of the group. Uniqueness of normal forms is obtained by using particular coset representatives for the passage from one copy of the graph of G to another.

- If a group acts geometrically on a geometry, how uniquely is the geometry determined? (Answer: the geometry is unique up to quasi-isometry.)

Uniqueness of the geometry on which a group acts geometrically. We note that the free Abelian group $Z \oplus Z$ on two generators acts geometrically on both its own Cayley graph and on the Euclidean plane. It is clear that these two geometries are not homeomorphic. However, if one stands far back, then the Cayley graph, which is the unit lattice in the plane, *looks* like the plane, just as a screen door looks like a door. That is, in the large the two geometries look alike. We can make this mathematical relationship precise by introducing the notion of *quasi-isometry*. Two spaces are quasi-isometric if they are Lipschitz equivalent in the large; local differences are ignored. Here are the precise definitions which we take from the discussion in the appendix of [13].

DEFINITION. A *relation* (multi-valued function) $R : X \to Y$ between spaces X and Y is said to be *quasi-Lipschitz* if

(1) R is everywhere defined (i.e., $R(x) \neq \emptyset$ for each $x \in X$); and

(2) there exist positive numbers, K and L, such that for each $A \subset X$

$$\operatorname{Diam} R(A) \leqslant K \cdot \operatorname{Diam} A + L.$$

'Quasi' (= to some extent), as we use it here, is simply a substitute for 'in the large.' If the summand L were omitted in (2), we would have the standard definition of Lipschitz. The summand L simply makes the inequality true for all small sets A provided only that R does not take small sets to arbitrarily large sets.

Some readers may feel uneasy in using relations instead of functions. We personally find many of the definitions more natural when one thinks in terms of relations. The reader who wants to insist on functions may use them, but these functions must be allowed to be discontinuous, since one is finding an equivalence relation between spaces which are not homeomorphic. If one uses relations, then one can use continuous relations. But that seems to be no conceptual advance. The functions and relations used are continuous in the large, whatever that means. Similarly, the spaces we shall look at in Section 4.1 are negatively curved in the large, though not necessarily negatively curved in the small. One really needs to be near-sighted to get the right point of view in these matters: with glasses off, local things blur and become inessential; the critical thing is the behavior in the large; do things behave in the large as they would if the property in question were satisfied in the small?

DEFINITION. Relations $R : X \to Y$ and $S : Y \to X$ are *quasi-inverses* if they are everywhere defined and there is a constant $M > 0$ such that $d(S \circ R, \mathrm{id}_X) < M$ and $d(R \circ S, \mathrm{id}_Y) < M$.

DEFINITION. A relation $R : X \to Y$ is *a quasi-Lipschitz equivalence* if there is a quasi-inverse $S : Y \to X$ such that both R and S are quasi-Lipschitz. If two spaces are quasi-Lipschitz equivalent, then we say that they are *quasi-isometric*.

EQUIVALENCE THEOREM. *If a group G acts geometrically on two geometries X and Y, then X and Y are quasi-isometric.*

PROOF. The proof is essentially an exercise. For complete details we refer the reader to the reference cited above. However, we do define suitable relations R and S.

Fix $\varepsilon > 0$. Let U_0 and V_0 be bounded, connected open sets in X and Y, respectively, such that the G-translates of U_0 and V_0 cover X and Y, respectively. Let U and V denote the ε-neighborhoods of U_0 and V_0, respectively. Define $R : X \to Y$ and $S : Y \to X$ by

$$R(x) = \{y \in Y \mid \exists g \in G,\ x \in gU,\ y \in gV\} \quad \text{and}$$
$$S(y) = \{x \in X \mid \exists g \in G,\ x \in gU,\ y \in gV\}.$$

Note that, as relations, $S = R^{-1}$. We claim that R and S are quasi-Lipschitz quasi-inverses. Distances are approximated by counting the number of translates of U or of V that are crossed in going from one point to another. □

3. Examples

Marshall Hall has been credited with the statement, "Theorems change, but the examples remain." In Section 3.1 we give finitely presented groups that act geometrically or almost geometrically on standard geometries. The reader can gain considerable insight by attempting to construct the associated Cayley graphs; by the equivalence theorem at the end of Section 2 these graphs are quasi-isometric to their associated standard geometries; therefore these graphs serve as instructive discrete approximations to the geometries in question. We postpone the description of the geometries to Section 3.2. In Section 3.3, we describe some of the Cayley graphs of these groups and show how the properties of the associated geometries manifest themselves in the combinatorics of the groups.

3.1. *Examples of finitely presented groups*

We attempt to describe the groups by a representative variety of means, and we choose the groups from a representative family of geometries. To the unpractised eye, these groups may look much alike. The point of the list is that they are associated naturally with a large variety of geometries and that they can be fruitfully studied by studying the appropriate geometries. The geometries will be described in Section 3.2.

3.1.1. As our first class of examples we take the free Abelian groups on a finite number n of generators. They have a finite presentation of the form

$$\langle x_1, \ldots, x_n \mid x_i x_j = x_j x_i \rangle,$$

or, using the commutator notation $[x, y] = xyx^{-1}y^{-1}$,

$$\langle x_1, \ldots, x_n \mid [x_i, x_j] = 1 \rangle.$$

As our second class of examples we take the Fibonacci groups originally defined by John Conway. A number of the relevant references occur in Lyndon and Schupp's book on combinatorial group theory [32, p. 97].

The natural presentation from which the groups received their name are these.

$$F_n = \langle x_1, \ldots, x_n \mid x_1 x_2 = x_3, \ldots, x_{n-1} x_n = x_1,\ x_n x_1 = x_2 \rangle.$$

All of these groups are interesting, but we are particularly interested in F_6.

Finally we add the group with presentation,

$$G = \langle a, b, t, A, C \mid ab = ba,\ t^2 = 1,\ tat = b,\ tbt = a,\ A = a^2,\ C = ab \rangle.$$

3.1.2 and 3.1.3. *Dodecahedra and pentagons.* The Poincaré dodecahedral space is a 3-manifold not homeomorphic with the 3-sphere but whose homology is isomorphic with that of the 3-sphere. One can define the space as an identification of the solid dodecahedron. One identifies opposite faces (in six pairs) by projecting a face through from one side of the dodecahedron to the other and then rotating the face just enough ($1/10$ of a turn clockwise) so that the two faces coincide. The fundamental group of this manifold is, according to Poincaré, not trivial. He discovered this example after making the initial false assertion that all homology 3-spheres are topological 3-spheres. He suggested the fundamental group as an invariant that would distinguish this particular 3-manifold from the 3-sphere and asked (the Poincaré conjecture) whether all simply connected, closed 3-manifolds are topological spheres.

The Seifert–Weber space is a 3-manifold that is formed again by identifying opposite faces of a dodecahedron, this time with $3/10$ of a turn.

The right-angled-dodecahedron reflection group is defined in terms of a regular dodecahedron in 3-dimensional non-Euclidean hyperbolic geometry. In that geometry, which will be described in Section 3.2, it is possible to find a regular dodecahedron all of whose dihedral angles are right angles. One then takes the group of non-Euclidean isometries generated by the twelve reflections in the faces of the dodecahedron.

There is a lower-dimensional analog of the group just described. Namely, in 2-dimensional non-Euclidean hyperbolic geometry there is a regular pentagon each of whose angles is a right angle. We may then take the group of hyperbolic isometries generated by the five reflections in the edges of this pentagon.

As groups analogous to the ones just described we include the fundamental groups of the closed, orientable surfaces:

$$G_g = \langle a_1, b_1, \dots, a_g, b_g \mid [a_1, b_1] \cdots [a_g, b_g] = 1 \rangle.$$

We use the notation $[a, b]$ for the commutator $aba^{-1}b^{-1}$.

3.1.4. *Knot groups.* Consider a polygonal knot (knotted circle) K in Euclidean 3-space R^3. Project the knot orthogonally into some plane P in such a fashion that any singular point of the projection is a double point where exactly two of the straight segments of K have crossing images where we think of one of the two edges as crossing under the other.

The image of K is a connected finite graph in the plane consisting of nonsingular arcs joining edge crossings. Form the dual graph by taking one vertex in each complementary domain and joining two of these dual vertices by an edge if the two complementary domains share a nonsingular boundary arc, one edge for each such boundary arc. One can realize each dual edge physically as an edge that starts at the dual vertex in one of the two domains, passes "under" the boundary edge to which it is dual and proceeds to the other dual vertex. If one orients the knot K, and hence the knot projection, then it is possible to orient the dual edges of the dual graph in such a way that, with the preferred orientation, the dual edge passes under its boundary edge from left to right. Now label the dual graph as follows.

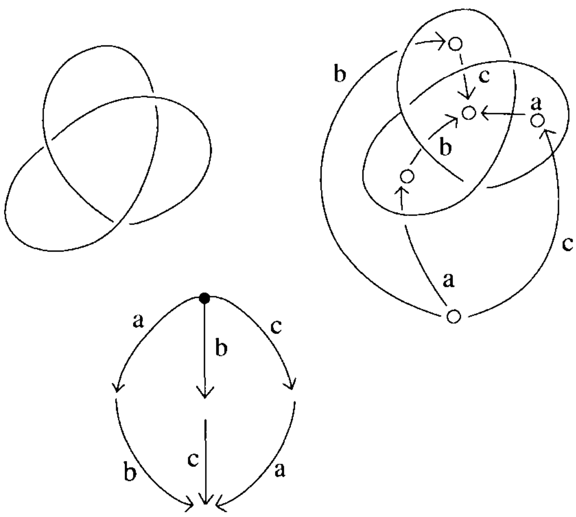


Fig. 2. The trefoil knot and its relator diagram.

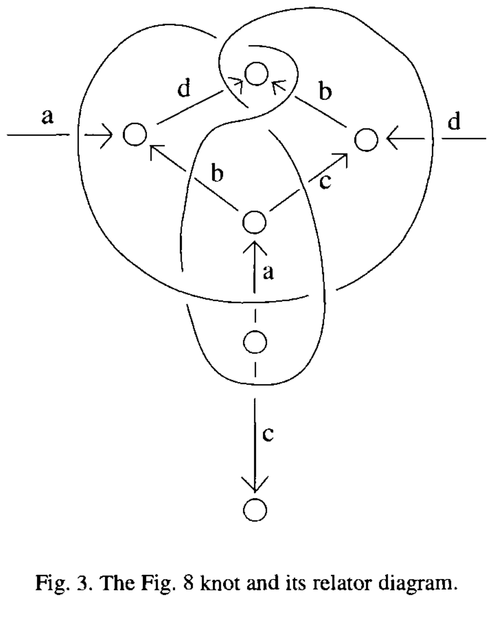


Fig. 3. The Fig. 8 knot and its relator diagram.

As one proceeds along the projection of K in the direction of its preferred orientation, one passes from one domain-boundary edge to another. Call two of these edges equivalent if the edge of K which connects them passes *over* its crossing edge. Extend to an equivalence relation on nonsingular boundary edges. Give two dual edges the same label provided that their corresponding boundary edges are equivalent. We call the labelled dual graph the *relator graph* of the knot projection. (Actually, the same procedure can be followed out if, instead of a knot K, we take a link in R^3.) Associated with the relator graph we have

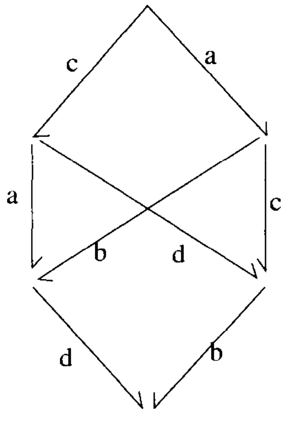


Fig. 4.

a group presentation. We take one generator for each label. We take one relator for each word labelling a closed edge path in the relator graph. All of the relators are actually consequences of any set of relators that form a homology basis for the 1-dimensional homology of the relator graph. We give as examples the trefoil knot group and the figure eight knot group. These presentations are presentations for the fundamental group of the complement of the knot.

3.1.5. We recall the soccer-ball group from the previous section:

$$\langle x, y \mid x^2 = y^5 = (xy)^3 = 1 \rangle.$$

We also consider the direct product of this group with the integers Z.

3.1.6. We consider the group defined by the presentation,

$$P = \langle x_1, y_1, x_2, y_2, t \mid t = [x_1, y_1][x_2, y_2],\ t \text{ central} \rangle.$$

3.1.7. We consider the matrix group N which consists of 3 by 3 matrices which are 1 on the diagonal, 0 below the diagonal, and integral above the diagonal.

3.1.8. We consider the group S given by the presentation

$$\langle a, b, t \mid ab = ba,\ tat^{-1} = b,\ tbt^{-1} = ab \rangle.$$

3.2. *The standard geometries in dimensions 2 and 3*

A good reference is [44].

As noted before, a geometry is *standard* if it is a simply connected, complete Riemannian manifold on which its own isometry group acts transitively on which some group can act geometrically and without fixed points. There is one standard geometry in dimension 1, namely the real line R. There are three standard geometries in dimension 2, namely the Euclidean plane, the 2-dimensional Euclidean sphere S^2, and the non-Euclidean hyperbolic plane H^2. Each of these geometries has an analogue in all dimensions greater than 2. Thurston proved that there are exactly eight standard geometries in dimension 3. We shall enumerate the eight 3-dimensional geometries in this section. For some of these geometries we will give a generic description that applies in dimensions greater than 3.

3.2.1. Euclidean n-space R^n has as its points the n-tuples $x = (x_1, \ldots, x_n)$ of real numbers and as its Riemannian metric $\mathrm{d}s^2 = \mathrm{d}x_1^2 + \cdots + \mathrm{d}x_n^2$.

3.2.2. The Euclidean n-sphere S^n has as its points the vectors $x = (x_1, \ldots, x_n, x_{n+1})$ from Euclidean $(n + 1)$-space which have Euclidean length 1 and as its Riemannian metric the restriction of the Euclidean metric on R^{n+1} to the tangent space of S^n.

3.2.3. Non-Euclidean hyperbolic n-space is, of all the standard geometries, by far the most important. Non-Euclidean hyperbolic n-space has a number of standard Riemannian models. A reasonably good short introduction to the properties of this space appears in the paper [16]. Other good references are [5,31,41].

We give here just one of the five standard models for hyperbolic n-space, namely Poincaré's upper half space model. As underlying space we take

$$H = \{(x_1, \ldots, x_n) \in R^n \mid x_n > 0\}.$$

As Riemannian metric we take

$$\mathrm{d}s^2 = (\mathrm{d}x_1^2 + \cdots + \mathrm{d}x_n^2)/x_n^2.$$

3.2.4 and 3.2.5. Euclidean geometry is, of course, a product geometry. There are two other standard product geometries in dimension 3. The first is the product $S^2 \times R$ of the 2-sphere with the line. The second is the product $H^2 \times R$ of the hyperbolic 2-dimensional plane with the line. In each case, one takes as the new Riemannian metric

$$\mathrm{d}s^2 = \mathrm{d}s_1^2 + \mathrm{d}s_2^2,$$

where $\mathrm{d}s_1^2$ is the standard metric on the first factor, $\mathrm{d}s_2^2$ the standard metric on the second factor.

3.2.6. There are three twisted geometries in dimension 3. The first of these is SL_2, or *the universal cover of the unit tangent bundle to the hyperbolic plane.* If we think of the hyperbolic plane in the upper half plane model, then its group of orientation preserving isometries is the group of linear fractional transformations with real coefficients:

$$z \mapsto \frac{az + b}{cz + d},$$

with a, b, c, and d real, and with $ad - bc = 1$. That is, this isometry group is $\mathrm{PSL}_2(R)$. This group of hyperbolic isometries induces homeomorphisms of the unit tangent bundle. One can extend this group of homeomorphisms by allowing, in addition, rotation of all fibers through some fixed real angle. These rotations commute with the other homeomorphisms we have described and commute as well with one another. That is, we obtain a group of homeomorphisms which has these fiber rotations in the center. After choice of base point, all of these homeomorphisms lift naturally to the universal cover SL_2. While in the base space, one cannot distinguish fiber rotations which differ by multiples of 2π, these homeomorphisms are naturally distinguished in the universal cover since the universal cover unwraps all of the circle fibers into lines. If one chooses the standard Euclidean Riemannian metric at the base point of SL_2, then the group of homeomorphisms we have described can be used to pull this metric back to all other points of SL_2. We obtain thereby a left invariant metric on SL_2.

This SL_2 is not isometric with $H^2 \times R$ but it is quasi-isometric with $H^2 \times R$.

3.7. The second of our twisted geometries is **Nil**. As a space it coincides with the space of 3 by 3 real matrices which are 0 below the diagonal, 1 on the diagonal, and arbitrary entries a, b, and c above the diagonal. This space is naturally homeomorphic with 3-dimensional Euclidean space with vectors (a, b, c). Matrix multiplication defines a group structure on this space.

$$(a, b, c) \cdot (\alpha, \beta, \gamma) = (a + \alpha, b + a\gamma + \beta, c + \gamma).$$

If one chooses the standard Euclidean Riemannian metric at the vector $(0, 0, 0)$, then the group structure may be used to pull the metric back to a left invariant metric at each point. The result is this:

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + (\mathrm{d}z - x\,\mathrm{d}y)^2$$

at the point (x, y, z).

3.8. The third and final twisted geometry is **Sol**. As a space it again coincides with Euclidean 3-space, the set of ordered triples (x, y, z) with the product topology. A group structure is defined on **Sol** by the formula

$$(a, b, c) \cdot (\alpha, \beta, \gamma) = \big(a + \alpha \cdot \mathbf{exp}(-c), b + \beta \cdot \mathbf{exp}(c), c + \gamma\big).$$

If one chooses the standard Euclidean Riemannian metric at the vector $(0, 0, 0)$, then the group structure may be used to pull the metric back to a left invariant metric at each point. The result is this:

$$\mathrm{d}s^2 = \mathbf{exp}(2z)\,\mathrm{d}x^2 + \mathbf{exp}(-2z)\,\mathrm{d}y^2 + \mathrm{d}z^2$$

at the point (x, y, z).

3.3. *The Cayley graphs of the examples and the geometries with which they are associated*

One begins to understand both the sample groups and the geometries with which they are associated as one constructs the Cayley graphs. The construction was left to the reader in Section 3.1; but now we describe the graphs for those who want to check their constructions or who failed to succeed in constructing them.

Before we begin, we give an easy example of how the nature of the geometries allows one to prove a simple group theoretic statement:

In reference to Gromov's polynomial growth theorem, we have observed the difference in growth rates between the free Abelian group Z^2 on two generators and the free non-Abelian group $Z * Z$ on two generators. We return to these simple examples. It is obvious the $Z * Z$ does not embed in Z^2 since the former is non-Abelian, the latter Abelian. But $Z * Z$ does embed in all of the fundamental groups of surfaces having negative Euler characteristic. What about the fundamental group K of the Klein bottle which has 0 Euler characteristic? The group K is non-Abelian. Does it contain an embedded copy of $Z * Z$. As one of the simplest sample applications of geometry to group theory that we know, we use geometry to prove that $Z * Z$ does not embed in K.

PROPOSITION. *The fundamental group K of the Klein bottle does not contain an embedded copy of the free non-Abelian group $Z * Z$.*

PROOF. The group K has presentation

$$K = \langle x, y \mid xyx^{-1}y = 1 \rangle.$$

It is easy to see from the recognition theorem that the Cayley graph of K may be realized as the unit square lattice in the complex plane with all horizontal edges pointing to the right being labelled x, vertical edges pointing upward over even integer values being labelled y, vertical edges pointing downward over odd integer values being labelled y.

Thus, as in the case of Z^2, K has exactly $4n$ elements of minimal length n for all integer lengths $n > 0$. If $Z * Z$ could be embedded in K, then the images of the generators would have fixed lengths bounded by some integer N. Thus the embedding of $Z * Z$ would induce an embedding of the elements of length $\leqslant n$ in $Z * Z$ into the elements of length $\leqslant N \cdot n$ in K. But there are $1 + 4 + 3 \cdot 4 + 3^2 \cdot 4 + \cdots + 3^{n-1} \cdot 4$ of the former but only $1 + 4 + 2 \cdot 4 + 3 \cdot 4 + \cdots + n \cdot N \cdot 4$ of the latter. The exponential sum of the former, however, clearly surpasses the quadratic sum of the latter for large n, a contradiction.

The reader will have surmised that the group examples from Section 3.1 and the standard geometries of Section 3.2 have been presented in parallel so that the groups act on the corresponding geometries. We now describe the Cayley graphs. □

3.3.1. *Euclidean groups.* The Cayley graph of the free Abelian groups Z^n can be realized as the unit lattice in Euclidean n-space R^n. The group acts geometrically on that space by translation in the direction of the standard unit vectors.

The Fibonacci groups are so named because of the nature of the relators. The relators can be used to reduce the generating set to two elements. One loses thereby much of the symmetry of the presentation. Conway's expressed goal in defining the groups was to create presentations for his students giving fairly simple groups in such a way that it was not at all obvious just how simple the groups were. In particular, F_5 is the cyclic group of order 11, but the proof that this is so requires consideration of huge powers arising from repeated exponentiation as one moves cyclically through the defining relators. The group F_6 acts geometrically on Euclidean 3-dimensional space. It is probably the only Euclidean group in the sequence. The group F_7 has order less than 30. All of the groups F_n for n large are infinite.

Some questions that have not been adequately dealt with are the following. Given natural infinite families of presentations, can one deal with most of the groups of the family in a uniform way? Are most of them infinite? Are most of them negatively curved (word-hyperbolic) in the sense of Gromov? (See Section 4.)

3.3.2. *Spherical groups.* Poincaré's dodecahedral space is finitely covered by the Euclidean sphere S^3. The dodecahedron lifts homeomorphically to S^3 and the 120 different lifts tile S^3 and are equivalent under the covering translations which are Euclidean rotations of S^3. One can build the universal cover and recognize it topologically as S^3 in a manner equivalent to Todd–Coxeter coset enumeration. One first calculates that three dodecahedra meet along each edge. One can view the patterns of three dodecahedra as defining relators for the fundamental group or as giving the necessary local conditions for the building of a covering space. One then begins with one dodecahedron and begins to build the universal cover by sewing new disjoint copies of the fundamental dodecahedron along free faces labelled according to opposite face identifications which defined the dodecahedral space. One satisfies the relators by making sure to put only three dodecahedra around each edge. If one does the construction systematically, say by forming at each stage the star of the previous stage, then one obtains successively larger balls whose boundary patterns are beautiful and easily calculated. At the final stage, a single final dodecahedron fills the hole and completes the sphere. It is very satisfying to carry out the construction and to see the sphere inevitably close up after the conjoining of 120 dodecahedra.

The examples that we have given of groups that act geometrically on the sphere have been particularly chosen so as to be obviously geometrically related to a sphere of some dimension. However, every finite group acts geometrically on every sphere: simply take the trivial action.

3.3.3. *Hyperbolic groups.* If the Poincaré dodecahedral space is modified just a tiny bit, the group becomes hyperbolic in the sense that it acts geometrically on hyperbolic 3-space; the Seifert–Weber space is in fact hyperbolic, not a finite group, but infinite.

While the torus group Z^2 and the Klein bottle group K are Euclidean, the fundamental groups of those surfaces that have negative Euler characteristic are hyperbolic. They present particularly nice examples of groups whose Cayley graphs, though infinite, can be quickly constructed algorithmically. As an example, we take the fundamental group of the closed, orientable surface of genus 2. In the standard way, we cut this surface along a bouquet of four loops so as to create a topological octagon with labelling boundary word $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1}$.

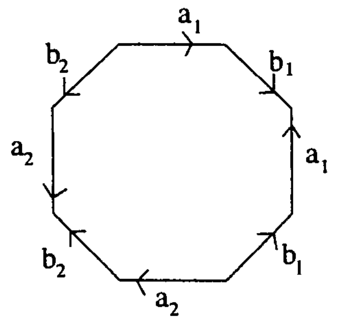


Fig. 5. Presentation for the fundamental group of the surface of genus two.

The universal covering space of the surface is sewn together from infinitely many disjoint copies of this labelled octagon. We begin with one such octagon. It forms a disk with boundary label. We also attach labels of 1 to each of the vertices to indicate that, in this disk, there is only one octagon adjacent to this vertex at the present time. We now proceed by induction. We assume inductively that we have a disk with boundary label sewn together from finitely many disjoint copies of our original labelled octagon. We assume that, in addition to the boundary edge labels, there are labels of 1, 3, 5, or 7 attached to the vertices of this disk which indicate how many octagons are already adjacent to that vertex. We assume further that each of the labels of 3 or 5 or 7 are in fact separated by at least one 1 (in fact by at least five 1's).

We form the next disk from this old disk by sewing new copies of the octagon along the boundary edges of our old disk, one disk along each edge *except* when we come to a vertex labelled with a 7. If a vertex is labelled with a 7, then, in order to form a covering space, a space locally homeomorphic to our original surface at that vertex, we are only allowed one more octagon. That final octagon must be sewn along the two adjacent edges which are joined at vertex labelled 7 to give eight octagons at that vertex. It is easy to see inductively that our new labelled disk satisfies the required inductive hypothesis. By a theorem entirely analogous to the recognition theorem, or by a suitable application of the recognition theorem to the fundamental group of the surface, it is easy to see that we construct larger and larger labelled disks whose infinite union is the universal cover of the surface. The reader should carry out several stages of the process in order to see just how compelling the construction is.

The readers who have successfully constructed the universal cover of the surface in the previous paragraph might test their skills with the construction of the tilings of the hyperbolic plane and of hyperbolic 3-space associated with the reflection groups in the right-angled pentagon and the right-angled dodecahedron. These groups have fixed points, and so they do not define a 2-manifold or a 3-manifold in the way that a geometric action without fixed points would do. However, it is easy to write down finite presentations for these groups by applying a very important theorem of Poincaré, called Poincaré's theorem for fundamental polygons and polyhedra. A good reference to the theorem is [33].

Maskit's work in general gives a very accessible approach to the study of hyperbolic groups. See, for example, [34].

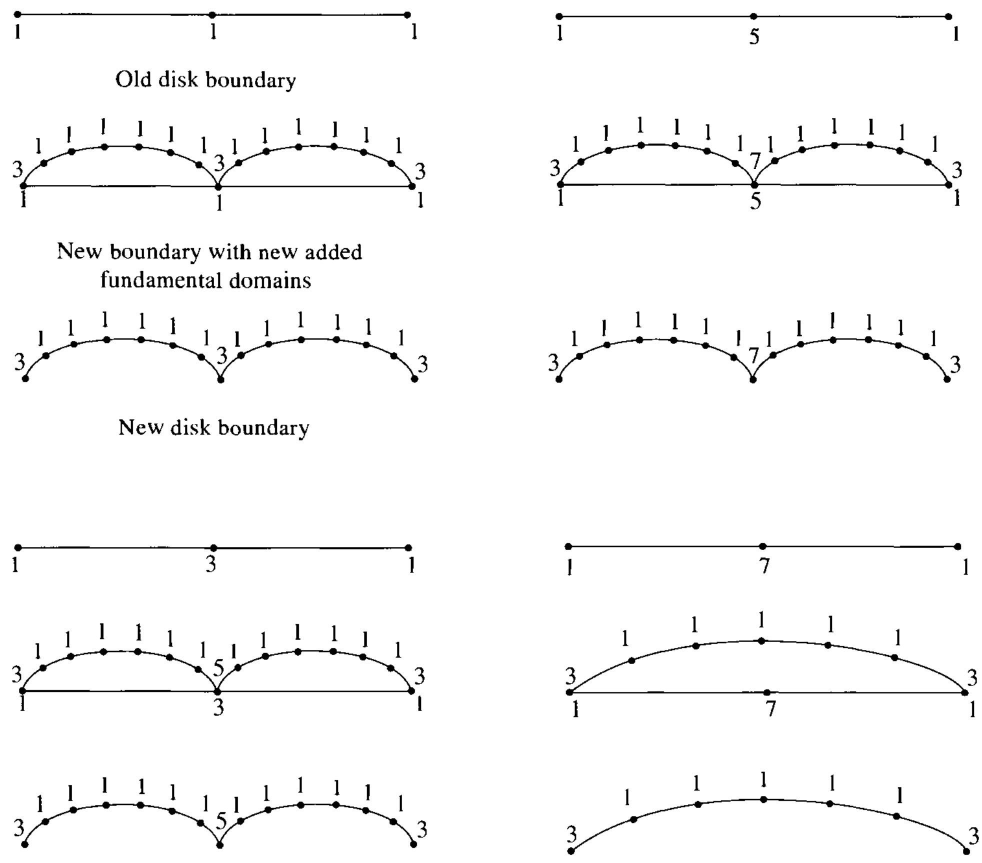


Fig. 6. Building the universal cover of the surface of genus two recursively. The four fundamental local patterns.

One should also see the work of Robert Riley [42] to see some rather fine applications of this theorem to 3-manifolds.

In our setting, Poincaré's theorem implies that one can construct something like a universal covering space by sewing together disjoint copies of the pentagon or dodecahedron in the freest possible fashion subject to the requirements that exactly four pentagons come together at each vertex and exactly four dodecahedra come together around each edge. In the pentagon case, one constructs labelled disks for which the vertices are all labelled 1 or 3, vertices labelled 3 separated by at least two vertices labelled 1.

The construction is then exactly like the construction of the universal cover of the surface of genus 2, except that there are fewer local patterns to recognize.

In the dodecahedron case, if one builds at each stage the complete star of the previous stage, then one finds recursively that every stage is a 3-cell and that the boundaries of the 3-cells are labelled in the following recursive fashion.

Thus, although one cannot actually carry out the infinitely many steps required to form the entire infinite pattern, one can carry out as many of the finite stages of the construction as time and space permit, and the recursion required for going from one stage to the next is completely clear.

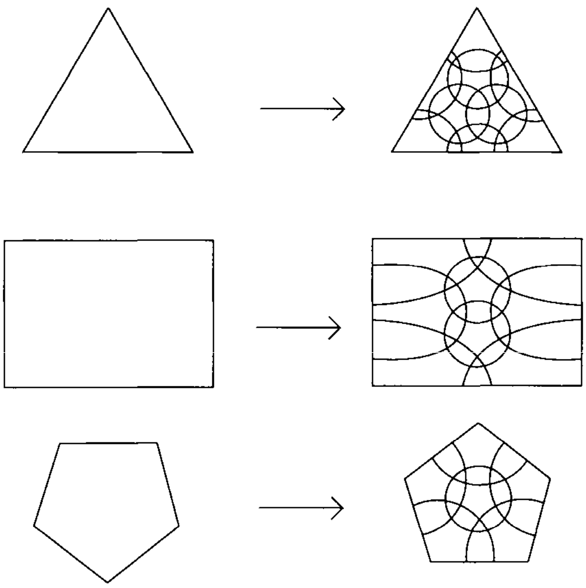


Fig. 7. Building the universal covering orbifold of the right-angled dodecahedral reflection group recursively. The three fundamental local boundary patterns.

3.3.4. The group defined in Section 3.1.4 is obviously associated with the very simple product geometry $S^2 \times R$. The soccerball factor is of course included only for aesthetic reasons. Since the factor S^2 is compact, it may be entirely ignored up to quasi-isometry. A group acts geometrically on $S^2 \times R$ if and only if it acts geometrically on the line R itself. The groups that so act are determined by Stallings's theorem on the ends of groups.

3.3.5. To obtain a group that acts geometrically on $H^2 \times R$, one may take the product of a group that acts geometrically on H^2 and multiply it by Z.

The free non-Abelian group $Z * Z$ does not act geometrically on H^2. However it does act isometrically and properly discontinuously on H^2 in such a way that the quotient, while not being compact, has finite area. The fundamental group G of the trefoil knot complement is almost the product of $Z * Z$ with Z. The commutator subgroup G' of G is the fundamental group of the incompressible, orientable Seifert surface of the knot, which is a torus with one hole, and which therefore has fundamental group $Z * Z$. The center C of the group G is isomorphic with Z. The internal product of G' and C in G is isomorphic with the external product of the same groups and has finite index in G. Thus, if we form the Cayley graph of G, we will have a reasonable approximation to the geometry $H^2 \times R$.

One of many ways to form the Cayley graph is as follows. In Euclidean xyz-space, form an infinite homogeneous tree in the plane $z = 0$, three edges meeting at each vertex, one vertex at the origin. Label each vertex of the tree alternately as either up or down, with the origin being up. Label each edge of the tree as red, white, or blue in such a manner that at each vertex there are edges of each color.

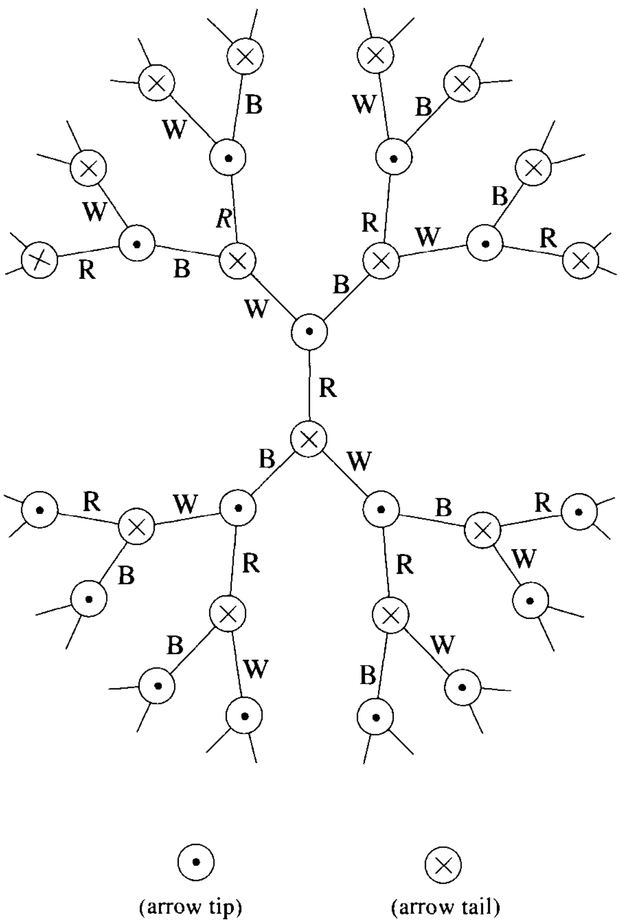


Fig. 8. One horizontal layer of the Cayley graph of the fundamental group of the complement of the trefoil knot.

Now take a copy of this graph in each of the planes z = integer, but change the labelling in the following way. As one moves up or down from one level to an adjacent one, change all up vertices to down vertices, all down vertices to up vertices. As one moves up from one level to an adjacent level, change red edges to blue edges, blue edges to white edges, white edges to red edges. Conversely, as one moves down from one level to an adjacent level, change red to white, white to blue, blue to red. Finally, join all of the levels together to form a single graph in the following way. Move all up vertices up $1/2$ unit. Move all down vertices down $1/2$ unit. Direct all edges so that they begin at an up vertex and end at a down vertex.

We recommend that the reader use the recognition theorem to show that the graph just described is the Cayley graph of the trefoil knot group. Here are some things that are easy to read from the Cayley graph. (1) There is a natural homomorphism onto Z; indeed, just collapse the relator diagram so that every directed edge points vertically downward, or collapse the entire graph to the vertical straight line through the origin with each edge of the graph going to a vertical interval of length 1. (2) The group is non-Abelian; simply follow paths labelled by two different generators in the two possible orders. (3) The element with labels red-white-blue-red-white-blue lies in the center of the group; indeed, check that it commutes with the three generators. (4) The commutator subgroup is the set of vertices lying in the plane $z = 0$.

We note that other knot groups are associated, in general, with different geometries.

3.3.6. We wish to describe the Cayley graph for the group given in Section 3.1.6.

$$S = \langle x_1, y_1, x_2, y_2, t \mid t = [x_1, y_1][x_2, y_2] \text{ central}\rangle.$$

It is clearly closely associated with the surface group

$$S_0 = \langle x_1, y_1, x_2, y_2 \mid [x_1, y_1][x_2, y_2] = 1\rangle,$$

whose Cayley graph we described above in Section 3.3.3. To form the Cayley graph of S we proceed as follows. We assume that the Cayley graph Γ_0 of S_0 has already been constructed and naturally embedded in the plane $z = 0$. (The most natural plane in which to embed it would be the hyperbolic plane, though that is not critical for our purposes.) We first concentrate only on the vertices. We take copies of the vertices in each plane $z =$ integer. We next take a tree T in Γ_0 which contains every vertex of Γ_0. Now to each edge path P in Γ_0 which begins at the identity vertex we assign a height $n(P)$ as follows. We consider the path P_0 in T which begins at the terminal endpoint of P and ends at the identity vertex of Γ_0. The closed path $P \cdot P_0$ encloses $n(P)$ of the disks in $z = 0$ bounded by the graph Γ_0, where a disk is counted k times if the path $P \cdot P_0$ has winding number k with respect to a point of the interior of the disk. The path P represents an element of the group S, namely that vertex which lies in the plane $z = n(P)$ above or below the terminal point of P which lies in the plane $z = 0$. If paths P and P' differ by one generating letter, say $P' = P \cdot x$, where x is either x_1, y_1, x_2, y_2, or their inverses, then their vertices are to be joined by an edge labelled by the generator x. Finally, each vertex is to be joined to the vertex directly above it by an edge labelled t. This completes the construction of the Cayley graph. This graph with each edge assigned the length of 1 approximates the geometry SL_2.

3.3.7. We shall see that the nil group of Section 3.1.7 has Cayley graph which can be constructed much in the fashion of the SL_2 group of Section 3.1.6 (see the previous Section 3.3.6) or alternatively much in the fashion of the solv group of Section 3.1.8 (see the next Section 3.3.8). We recall the description. Our nilgroup N is a matrix group consisting of those 3×3-matrices of the form

$$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}.$$

Define the three matrices

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \qquad B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \qquad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Easy calculations establish the following relationships:

$$C^c A^a B^b = \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix};$$

$$ACA^{-1}C^{-1} = B;$$

$$AB = BA;$$

$$CB = BC.$$

The first of the relationships shows that the three elements A, B, and C generate the group. The last two relators show that B lies in the center of the group, and an easy calculation shows that in fact B generates the center.

We claim that in fact this matrix group has as presentation

$$\langle A, B, C \mid [A, C] = B,\ [A, B] = 1,\ [C, B] = 1 \rangle,$$

where $[X, Y]$ denotes the commutator $XYX^{-1}Y^{-1}$. The group given by this presentation clearly can be mapped homomorphically into the matrix group since the matrix group satisfies the three relators. The map is onto by the first relationship above. The three relators can clearly be used to put any element into the form $C^c A^a B^b$, and these forms map one to one into the matrix group so that the homomorphism is injective.

We may now copy the construction of the Cayley graph from the previous section: we start with a surface group (a free-Abelian torus group $Z \oplus Z$ generated by A and C); we then delete the relator that says that A and C commute and replace their commutator instead by a new generator B (analogous to the generator t of the previous section). We then require that this new generator be central.

There is an alternative way of constructing the Cayley graph which uses a little bit of linear algebra. We shall use an identical construction in the next Section 3.3.8.

In order to form the Cayley graph, we first modify the presentation by the substitutions of generators $\alpha = A^{-1}$, $\beta = B$, and $t = C$. The presentation then takes the form

$$\langle \alpha, \beta, t \mid [\alpha, \beta] = 1,\ t\alpha t^{-1} = \alpha\beta,\ t\beta t^{-1} = \beta \rangle.$$

We use the matrix

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^n.$$

Here is the construction: In each integral z-level put a copy of the unit lattice, the graph of the free Abelian group $\mathbf{Z}^2$ generated by α and β. Move the lattice in z-level n by the linear mapping

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^n.$$

Thus, the further one moves away from the level $z = 0$, the further the lattice deviates from the square lattice. Since the matrix is invertible over the integers, the vertices of the different lattices are still the integer points. Hence we may join each vertex to the one above it by an edge labelled t. The Cayley graph is now the union of the distorted integer lattices and the edges labelled t.

Though the form of the construction is identical with that of the next section, the associated geometries are different because the monodromy matrix associated with conjugation by t is not hyperbolic but consists instead of a single Jordan block with multiple eigenvalues equal to 1.

3.3.8. Consider the group with the following presentation:

$$\langle a, b, t \mid aba^{-1}b^{-1},\ tat^{-1} = ab,\ tbt^{-1} = ab^2 \rangle.$$

In each integral z-level put a copy of the unit lattice, the graph of the free Abelian group $\mathbf{Z}^2$ generated by a and b. Move the lattice in z-level n by the linear mapping

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^n.$$

Thus, the further one moves away from the level $z = 0$, the further the lattice deviates from the square lattice. For n large and positive, every edge seems to point more and more in the eigendirection of the matrix associated with the eigenvalue greater than 1. For n large and negative, every edge seems to point more and more in the eigendirection of the matrix associated with the eigenvalue smaller than 1. However, since the matrix is invertible over the integers, the vertices of the different lattices are still the integer points. Hence we may join each vertex to the one above it by an edge labelled t. The Cayley graph is now the union of the distorted integer lattices and the edges labelled t.

4. Abstract classes of groups

The geometries and groups which we have explored in Section 3 have served as models for the development of geometric group theory over the last 20 years. The goal has been to develop combinatorial techniques to deal with all geometric 3-manifold groups and their generalizations. A great deal of success has been attained in at least three abstract classes of groups: (1) Gromov's word-hyperbolic groups which generalize the geometric properties of the groups which act geometrically on hyperbolic non-Euclidean geometry; (2) the automatic groups which distil the asymptotic recursive and computational properties noted by Cannon in his study of the same class of groups; and (3) the CAT(0) groups which attempt to reach out from groups of negative curvature to those of curvature $\leqslant 0$. In this section we give a short outline of these three subjects. In the past decade there have been a number of books and many, many papers written on these subjects. We cannot hope to duplicate them in this survey.

In the midst of this development, there has been a vast explosion of knowledge about geometric invariants of groups. We shall discuss some of these invariants in Section 6.

The whole development is somewhat reminiscent of the original development of combinatorial group theory originally following the early geometric work of Max Dehn. The early geometric insights became quite rapidly obscured as people turned to the more abstract and combinatorial realizations of those geometric ideas. That development led to small cancellation theory, which Gromov's work has once again made geometric rather than so exclusively combinatorial.

4.1. *Word hyperbolic or negatively curved groups*

The fundamental paper is [29].

The paper is truly an essay in the sense that it simply outlines a vast amount of work and, when it gives details at all, they are often sketchy or even questionable. Nevertheless, this is a tremendously important paper. Gromov has connected group theory with global differential geometry, and in that aspect of the work he has been *very* careful. As a consequence he is able to translate geometric theorems into group theory with great freedom. In this translation he has not been as careful, and so has left much work to be completed. Many, many group theorists and geometers have therefore been involved in sorting out the details. This fundamental paper should not be read linearly since the ideas necessary to complete the arguments of one section may appear in a widely separated section. It is easier, but not as rewarding, to read one of the several books dedicated to making it all work precisely. Here are a few of the expository treatments [1,6,19,27].

Two closely related papers of Cannon [11,13] may be helpful as regards some of the results.

Defining properties. Gromov isolated a whole family of equivalent properties of a metric space that are satisfied by manifolds having negative sectional curvature such as non-Euclidean hyperbolic space. Each is useful under appropriate circumstances, but the consensual favorite seems to be Rips's *thin triangles* property. We say that a proper path metric space X with path metric d has uniformly thin triangles (δ) if, whenever xyz is a geodesic triangle in X and whenever p is a point of one of the three sides, say $p \in xy$, then there is a point q in one of the other two sides, xz and yz, such that $d(p, q) \leqslant \delta$. For example, any tree has uniformly thin triangles (0). What is not so obvious is that non-Euclidean hyperbolic space H^n has uniformly thin triangles $(\log(1 + \sqrt{2})) = 0.88137\ldots$. As a consequence, geodesic polygons in such a space look essentially like slightly fattened trees. Thus geometric arguments become largely combinatorial in nature.

A group is said to be *word-hyperbolic* if its Cayley graph with respect to some finite generating set has uniformly thin triangles. A fundamental theorem shows that this property is independent of the finite generating set. Since the fundamental groups of closed manifolds with negative sectional curvatures are word-hyperbolic and since spaces that have uniformly thin triangles behave in the large just as negatively curved spaces behave in the large, the author also likes to call such spaces and groups *negatively curved* (in the large).

Genericity. As odd as the thin triangles property seems at first glance, Gromov proved first of all that it is very easy to construct large numbers of spaces that satisfy this property. Among these spaces are the Cayley graphs of many groups. But even more, Gromov claimed that the groups satisfying this property are in fact generic. A.Yu. Ol'shanskii [40] has confirmed that this is indeed the case: in a definite statistical sense, almost every finitely presented group is word-hyperbolic.

The word, conjugacy, and isomorphism problem s. In the paper by Cannon [11] referenced above, Cannon proved that groups which act geometrically on hyperbolic space have word problem which is very efficiently solvable by a local shortening procedure known as *Dehn's algorithm*; Dehn developed this method in order to solve the word problem for surface groups. Cannon's proof works essentially without change for word-hyperbolic groups. The existence of such an algorithm for solving the word problem for a group is in fact equivalent to the word-hyperbolicity of the group. A suitable reference is [25] (see in particular the second appendix).

Dehn also solved the *conjugacy problem* for surface groups. This problem asks for an algorithm which decides whether two words represent conjugate elements in the group. Cannon showed that Dehn's solution to that problem also extended to groups acting geometrically on hyperbolic space and the proof extends to word-hyperbolic groups. Dehn also gave an algorithm to determine when two surface groups are isomorphic. Zlil Sela has managed to solve the isomorphism problem for word-hyperbolic groups.

Gromov has given other simple procedures to solve the word problem in negatively curved groups.

Important unsolved problems. The following problems seem to be very difficult.

QUESTION. *Does every word-hyperbolic group have a torsion free subgroup of finite index?*

QUESTION. *Is every word-hyperbolic group residually finite?*

One of the fundamental properties of a word-hyperbolic group is that it has a well-defined space at infinity. This space at infinity intervenes at many places in theorems about word-hyperbolic groups.

QUESTION. *Suppose a word-hyperbolic group has as its space at infinity the* 2*-sphere* S^2. *Does it follow that the group acts geometrically on hyperbolic* 3*-space?*

The space at infinity. The points at infinity are equivalence classes of geodesic rays in the Cayley graph of the group. Two rays are equivalent if they remain asymptotically a finite distance apart. Two points at infinity are near to one another if representative geodesic rays remain close to one another for a long time. The space at infinity is compact, metrizable, finite-dimensional. It compactifies the Cayley graph in such a way that the group acts not only on the Cayley graph by left multiplication but also on the union of the graph with the space at infinity. It has just recently been proved by Brian Bowditch [7] and Gadde A. Swarup [46] that, if the space at infinity is connected, then it is also locally connected.

Nadia Benakli [4] has given examples of word-hyperbolic groups having the Sierpinski curve and the Menger curve as their respective spaces at infinity.

Here is an unresolved question.

QUESTION. *Characterize topologically those topological spaces that can occur as the space at infinity for a negatively curved group.*

4.2. *Automatic groups*

The author learned about Cayley graphs and Todd–Coxeter coset enumeration from the book by Coxeter and Moser [17].

Taking it as an extended exercise to construct the Cayley graph of all of the groups whose presentations were given in the book, the author discovered many beautiful graphs. One of the first was the graph of the trefoil knot group described in Section 3. Since the author had somehow been led to expect that it would be essentially impossible to understand the graphs of all except the simplest infinite groups, he was surprised to find again and again that the asymptotic structure of the Cayley graphs he considered were eminently understandable.

This discovery led to the paper cited above (Cannon [11]). Particularly Theorem 7 of that paper considered the combinatorial pictures as one looked toward infinity from the identity vertex through different vertices of the group. That theorem showed that there were in fact, for negatively curved groups, only finitely many such pictures. Once one had discovered these finitely many pictures, one knew what was going to happen *forever* in the Cayley graph.

As a consequence, one learned that the language of geodesic words in a negatively curved groups form what is called a *regular language*. Thurston was later to distil the essence of this situation to create *automatic group theory*. The basic reference for automatic group theory is [22].

We shall give a short introduction to automatic group theory here.

Finite state automata. We start with a finite alphabet A. We consider the set W of (finite) words that can be written in the alphabet A, including the empty word. We now consider a finite directed and labelled graph Λ with the following properties.

(1) From each vertex emanates one and only one edge labelled by each letter of the alphabet A. Consequently, beginning at any given vertex, and given any word $w \in W$, there is one and only one path labelled with the word w. It leads from the given vertex to another vertex of the graph Λ.

(2) One vertex of the graph Λ is designated as the *start vertex* v_0 of Λ.

(3) Some family S of vertices of Λ (possibly the empty subset) is designated as the family of *success vertices* or *accept vertices* of Λ.

The graph Λ with its start vertex and its accept vertices is called a *finite state automaton*. It defines a subset of W called its *language* $L(\Lambda)$ *of accepted words* in the following way. Take a word $w \in W$. Take the unique path which begins at the start vertex v_0 and is labelled by the word w. The word is *accepted* if the terminal vertex of the path is an accept or success vertex. The set of accepted words is the language $L(\Lambda)$.

A language is called *regular* if it is the language accepted by some finite state automaton. The set of regular languages has very nice set theoretic properties. It is closed under unions, intersections, complements, and various other nice logical operations. Such languages lend themselves very nicely to machine computation. The search operations in most word processing units are based on the theory of regular expressions or regular languages. This fact motivates the name of the reference, "Word processing in groups."

The fellow traveller property. It is convenient to parametrize an edge path in a Cayley graph by arc length. It is also convenient to pad such a finite path so that the path is parametrized by $[0, \infty)$; one simply has the path sit forever at the final vertex after it reaches that vertex. With those conveniences seen to, we can define the *k-fellow traveller property.* We say that two paths satisfy the k-fellow traveller property if, for each parameter $t \in [0, \infty)$, the image points in the two paths are within k of one another in the Cayley graph.

Definition of automatic group. A group G is said to be *automatic* if there is a finite alphabet A, a map $\phi : A \to G$, and a finite state automaton Λ on the alphabet A, and a positive number k having the following properties.

(1) First we note that the map ϕ can be extended to the entire set W of words written in the alphabet A by concatenation. The extended map $\phi : A \to G$ must map the regular language $L(\Lambda) \subset W$ onto G. That is, the elements of G must all be represented by accepted words of Λ.

(2) If u and v are accepted words of G and if the associated paths in the Cayley graph, which begin at the identity vertex and are labelled respectively by u and v, end within a distance 1 of each other, then these paths satisfy the k-fellow traveller property.

There are two key facts about automatic groups.

The first key fact is that in such groups one can perform computations very efficiently. David Epstein, Derek Holt, and Sarah Rees at Warwick University have developed an entire suite of computer programs designed to discover automatic structures on finitely presented groups when they exist and to create all of the finite state automata necessary to make quick and efficient calculations. Their suite of programs is called "aut," for *automatic groups*.

The second key fact is that such groups are generic in the class of all groups. In fact, the fundamental example from automatic group theory is that every word hyperbolic group is automatic. As the regular language required, one may take the language of geodesic words in the generating set of the group. Property (2) is then a reasonably easy consequence of the thin triangles property: geodesics in a negatively curved group which have their endpoints near to one another must stay at all times near to one another.

Using the set theoretic properties of regular languages, one can clean up the automata used. In particular, one can pass to an accepted language of unique representatives for the group G. That is, one can represent the elements of G by quickly calculable normal forms.

4.3. CAT(0) *groups*

As we noted in Section 4.1, negatively curved or word-hyperbolic groups form a generic class of finitely presented groups. But of course all 3-manifold groups are interesting to topologists, and the negatively curved groups certainly do not include the fundamental groups associated with five of the other standard 3-dimensional geometries (the groups associated with S^3 being finite are negatively curved in a trivial sense; the groups associated with $S^2 \times R$ are negatively curved since Z as a rather trivial tree is negatively curved). If geometry is to be used as a guiding principle in the study of groups, then one has to seek generalizations which pull in other geometries. Various generalizations have been considered. All of these have included in addition at least the Euclidean groups. The generalization to automatic groups introduced in the previous section is computationally based. The generalization of this section is geometrically based. Among the fundamental references are [2,9,10].

DEFINITION. Let X denote a geometry. If xyz is a geodesic triangle in X, then the sides satisfy the triangle inequality so that there is a *comparison triangle* $x'y'z'$ in the Euclidean plane which has identical edge lengths: $|xy| = |x'y'|$, $|yz| = |y'z'|$, and $|zx| = |z'x'|$. There is an obvious correspondence between the two triangles which is an isometry on corresponding edges. We say that X is a CAT(0) space if, for each choice of vertex v of xyz and for each choice of point w on the edge opposite v, $d(v, w) \leqslant d(v', w')$ where v' and w' are the points of the comparison triangle $x'y'z'$ corresponding to the points v and w of xyz. That is, in the space X triangles are at least as thin as the triangles in Euclidean space. The CAT(0) property is intended to generalize the notion of space with sectional curvatures $\leqslant 0$.

DEFINITION. We say that a group is a CAT(0) group if it acts geometrically on a CAT(0) space.

An interesting first example of a CAT(0) space that is neither negatively curved nor Euclidean is the Cartesian product of the Cayley graph of the free non-Abelian group on two generators and the Cayley graph of the integers. The first is an infinite tree in which four edges come together at each vertex. The second is the line. The product group $Z * Z \times Z$ acts geometrically on this CAT(0) space and is therefore a CAT(0) group.

The class of CAT(0) spaces shares many of the important characteristics of negatively curved spaces. In particular such spaces can be locally recognized so that there are nice constructions which stay in this class of spaces. Also the fundamental decision problems, the word and conjugacy problems, can be solved in this class of spaces, though their solutions are not as nice as they are for negatively curved groups. CAT(0) spaces have a natural "visual" space at infinity which can serve as space at infinity for CAT(0) groups; this visual space at infinity has as its set of points the geodesic rays emanating from some base point; two rays are close together if they remain close together over a large initial segment of each. Unfortunately, these spaces at infinity are not nicely attached to the groups in question. For example, with the CAT(0) group $Z * Z \times Z$ described in the previous paragraph, the space at infinity is the suspension of the Cantor set, and the group does not act continuously on the union of the Cayley graph and the space at infinity. The appropriate reference to this fact is [8].

5. Other geometries to explore

We have considered primarily the standard 2- and 3-dimensional geometries arising in the study of topological manifolds. However, there are many other geometries which might be studied as sources for appropriate geometric group theory. There are, of course, all of the classical Lie groups. One particularly relevant reference is [49].

One might study the affine geometries, the projective geometries, the conformal geometries. One might study the groups associated with important classes of Riemannian manifolds, such as the Kähler manifolds or Einstein manifolds.

In order to have a good theory, as pointed out by Gromov and Bridson, it is particularly important to be able to construct examples. It is often amazingly difficult to construct examples.

One should mention in passing that there are two essentially nongeometric sources of examples for combinatorial or geometric group theory, namely logic and algebra. Much work remains to be done in connecting these sources with geometry. A particularly interesting recent paper in this regard is the paper of Nabutovsky [39].

Nabutovsky gives "a lower bound for the number of contractible closed geodesics of length $\leqslant x$ on a compact Riemannian manifold M in terms of the resource-bounded Kolmogorov complexity of the word problem for $\pi_1(M)$, thus answering a question posed by Gromov."

6. Other geometric invariants

We mention here just a few of the geometric invariants that have been introduced and studied by geometric group theorists in the past decade. We apologize in advance for the many that we omit. We refer to possible entry papers into the topics cited rather than to the most recent or most complex papers on the subject. The most amazing collections of geometric invariants appear in the two papers of Gromov [29,30].

One of the very interesting invariants introduced by Gromov is that of isoperimetric inequalities: If w is a word representing the identity element in G, then w can be expressed as a product of conjugates of defining relators; this product may be thought of as defining a disk bounded by w and tiled by small subdiscs corresponding to the relators of the product; the number of small disks required is called the *area* of the disk. The *isoperimetric problem* asks for the smallest area of a disk bounded by w. A Dehn function $f(n)$ for the group is a function such that, if w is a relator of length $\leqslant n$, then w bounds a disk of area $\leqslant f(n)$. Gromov has characterized the word-hyperbolic or negatively curved groups as those which admit a linear Dehn function. Great progress has been made in calculating the Dehn functions and the *possible* Dehn functions of groups.

Very interesting results about isoperimetric functions on groups can be obtained by applying homological and cohomological methods. We mention two references [3,24].

Our personal attempt to generalize negatively curved groups so as to include the Euclidean groups involved the notion of *almost convex groups*. Here are two references [12,15].

Casson and Poenaru have suggested a generalization which has application to the following problem: when is the universal covering space of an irreducible, orientable 3-manifold homeomorphic with Euclidean 3-dimensional space R^3? The required condition C_2 is a property of the fundamental group of the manifold and is called an *isodiametric inequality*. It may be interpreted as a measure of how efficiently the Todd–Coxeter enumeration completes the correct construction of the ball of radius n in the Cayley graph of that group. If the process works efficiently enough, then the universal cover is, in fact Euclidean. An exposition of this work appears in the reference [26].

Much fruitful work has been expended in extending Stallings's work on the ends of groups. One very beautiful consequence is the accessibility theorem for finitely presented groups proved by Dunwoody [21].

Dunwoody proves that, in a finitely presented group, it is impossible to factor it nontrivially an infinite number of times as a free product with amalgamation or HNN extension.

Others have attempted to analyze the topology of the group near the ends of the group. Geoghegan and Mihalik have studied what they call *semistability* of the group at an end of the group. They have also studied the *fundamental group at infinity*, and other topological properties of the ends of a group. We give only two entry references [23,36].

References

[1] J.M. Alonso, T. Brady, D. Cooper, V. Ferlini, M. Lustig, M. Mihalik, M. Shapiro and H. Short, *Notes on word-hyperbolic groups*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific, Singapore (1991), 3–63.

[2] W. Ballmann, M. Gromov and V. Schroeder, *Manifolds of Nonpositive Curvature*, Birkhäuser, Boston (1985).

[3] G. Baumslag, C.F. Miller III and H. Short, *Isoperimetric inequalities and the homology of groups*, Invent. Math. **113** (1993), 531–560.

[4] N. Benakli, *Polygonal complexes. I. Combinatorial and geometric properties*, J. Pure Appl. Algebra **97** (1994), 247–263.

[5] R. Benedetti and C. Petronio, *Lectures on Hyperbolic Geometry*, Universitext, Springer, Berlin (1992).

[6] B.H. Bowditch, *Notes on Gromov's hyperbolicity criterion for path-metric spaces*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific, Singapore (1991), 64–167.

[7] B.H. Bowditch, *Connectedness properties of limit sets*, Trans. Amer. Math. Soc. **351** (1999), 3673–3686.

[8] P.L. Bowers and K. Ruane, *Boundaries of nonpositively curved groups of the form $G \times Z^n$*, Glasgow Math. J. **38** (1996), 177–189.

[9] M.R. Bridson, *Geodesics and curvature in metric simplicial complexes*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific, Singapore (1991), 373–463.

[10] M.R. Bridson and A. Haefliger, *Metric Spaces of Non-Positive Curvature*, Springer, Berlin (1999).

[11] J.W. Cannon, *The combinatorial structure of cocompact discrete hyperbolic groups*, Geom. Dedicata **16** (1984), 123–148.

[12] J.W. Cannon, *Almost convex groups*, Geom. Dedicata **22** (1987), 197–210.

[13] J.W. Cannon, *The theory of negatively curved spaces and groups*, Ergodic Theory, Symbolic Dynamics, and Hyperbolic Spaces, T. Bedford, M. Keane and C. Series, eds, Oxford Univ. Press, Oxford (1991), 315–369.

[14] J.W. Cannon, *Colored graphs*, Expository preprint, 27 pages.

[15] J.W. Cannon, W.J. Floyd, M.A. Grayson and W.P. Thurston, *Solvgroups are not almost convex*, Geom. Dedicata **31** (1989), 291–300.

[16] J.W. Cannon, W.J. Floyd, R. Kenyon and W.R. Parry, *Hyperbolic geometry*, Flavors of Geometry, S. Levy, ed., MSRI Publ., Vol. 31, Springer, Berlin (1997), 59–115.
[17] H.S.M. Coxeter and W.O.J. Moser, *Generators and Relations for Discrete Groups*, Ergebnisse der Mathematik, Bd. 14, 2nd ed., Springer, Berlin (1965).
[18] H.S.M. Coxeter and W.O.J. Moser, *Generators and Relations for Discrete Groups*, Ergebnisse der Mathematik, Bd. 14, Springer, Berlin (1972).
[19] M. Coornaert, T. Delzant and A. Papadopoulos, *Notes sur les groupes hyperboliques de Gromov*, Lecture Notes in Math., Vol. 1441, Springer, Berlin (1990).
[20] M. Dehn, *Über unendliche diskontinuierliche Gruppen*, Math. Ann. **71** (1912), 116–144.
[21] M.J. Dunwoody, *The accessibility of finitely presented groups*, Invent. Math. **81** (1985), 449–457.
[22] D.B.A. Epstein, with J.W. Cannon, D.F. Holt, S.V.F. Levy, M.S. Paterson and W.P. Thurston, *Word Processing in Groups*, Jones and Bartlett, Boston (1992).
[23] R. Geoghegan and M. Mihalik, *The fundamental group at infinity*, Topology **35** (1996), 655–669.
[24] S.M. Gersten, *Isoperimetric functions of groups and exotic cohomology*, Combinatorial and Geometric Group Theory (Edinburgh, 1993), London Math. Soc. Lecture Note Series, Vol. 204, Cambridge Univ. Press, Cambridge (1995), 87–104.
[25] S.M. Gersten and H.B. Short, *Small cancellation theory and automatic groups*, Invent. Math. **102** (1990), 305–334.
[26] S.M. Gersten and J.R. Stallings, *Casson's idea about 3-manifolds whose universal cover is* R^3, Internat. J. Algebra Comput. **1** (1991), 395–406.
[27] E. Ghys and P. de la Harpe, eds, *Sur les groupes hyperboliques d'aprés Mikhael Gromov*, Progr. Math., Vol. 83, Birkhäuser (1990).
[28] M. Gromov, *Groups of polynomial growth and expanding maps*, Publ. Math. IHES **53** (1981), 53–73.
[29] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, S.M. Gersten, ed., MSRI Publ., Vol. 8, Springer, Berlin (1987), 75–265.
[30] M. Gromov, *Asymptotic invariants of infinite groups*, Geometric Group Theory: Proceedings of the Symposium Held in Sussex, 1991, Cambridge Univ. Press, Cambridge (1993).
[31] B. Iverson, *Hyperbolic Geometry*, London Math. Soc. Student Texts, Vol. 25, Cambridge Univ. Press, Cambridge (1993).
[32] R.C. Lyndon and P.E. Schupp, *Combinatorial Group Theory*, Springer, Berlin (1977).
[33] B. Maskit, *On Poincaré's theorem for fundamental polygons*, Adv. in Math. **7** (1971), 219–230.
[34] B. Maskit, *Kleinian Groups*, Springer, Berlin (1988).
[35] W.S. Massey, *Algebraic Topology: An Introduction*, Springer, New York (1967).
[36] M. Mihalik, *Semistability at* ∞, ∞*-ended groups and group cohomology*, Trans. Amer. Math. Soc. **303** (1987), 479–485.
[37] J.W. Morgan and H. Bass, *The Smith Conjecture*, Academic Press, Orlando (1984).
[38] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Ann. of Math. Stud., Vol. 78, Princeton Univ. Press, Princeton, NJ (1973), 195.
[39] A. Nabutovsky, *Fundamental group and contractible closed geodesics*, Comm. Pure Appl. Math. **49** (1996), 1257–1270.
[40] A.Yu. Ol'shanskii, *Almost every group is hyperbolic*, Internat. J. Algebra Comput. **2** (1992), 1–17.
[41] J.G. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Graduate Texts in Math., Vol. 149, Springer, New York (1994).
[42] R. Riley, *Applications of a computer implementation of Poincaré's theorem on fundamental polyhedra*, Math. Comp. **40** (1983), 607–632.
[43] J.J. Rotman, *The Theory of Groups*, Allyn and Bacon, Boston (1973).
[44] P. Scott, *The geometries of 3-manifolds*, Bull. London Math. Soc. **15** (1983), 401–487.
[45] J.R. Stallings, *On torsion-free groups with infinitely many ends*, Ann. Math. **88** (1968), 312–334.
[46] G.A. Swarup, *On the cut point conjecture*, Electron. Res. Announc. Amer. Math. Soc. **2** (2) (1996), 98–100.
[47] W. Thurston, *The geometry and topology of 3-manifolds*, Lecture Notes, Dept. of Math., Princeton Univ. Press, Princeton, NJ (1978).
[48] W.P. Thurston, *Three-Dimensional Geometry and Topology*, Vol. 1, S. Levy, ed., Princeton Univ. Press, Princeton (1997).
[49] J.A. Wolf, *Spaces of Constant Curvature*, Publish or Perish, Inc., Berkeley, CA (1977).

CHAPTER 7

Infinite Dimensional Topology and Shape Theory

Alex Chigogidze*

Department of Mathematics and Statistics, University of Saskatchewan, McLean Hall, 106 Wiggins Road, Saskatoon, SK, S7N 5E6, Canada

E-mail: chigogid@math.usask.ca

Contents

---

*The author was partially supported by NSERC research grant.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

Infinite dimensional topology, in its traditional understanding, is a branch of geometric topology which studies infinite dimensional spaces, arising naturally in topology and functional analysis. Objects such as the Hilbert cube Q (i.e., the countable infinite product I^{ω} of the closed interval), the Hilbert space l_2 (topologically the countable infinite product R^{ω} of the real line), σ-compact locally convex linear spaces Σ and σ, as well as manifolds modeled on them, are typical examples of spaces attracting attention of experts in this field. Although the first nontrivial results in this direction (for example, the theorem of Keller [109] on the topological homogeneity of I^{ω}) appeared more than sixty years ago, systematic studies of infinite dimensional manifolds became possible only after discovery by K. Borsuk of ANE-spaces. Standard tools of the theory of ANE-spaces have their roots in classical methods of piecewise linear topology as well as in homological and homotopical methods of CW complexes. Long awaited topological characterizations of the above mentioned model spaces, as well as of the corresponding manifolds, were obtained by Toruńczyk [140, 141] in terms of certain strong universality conditions (see also [21, 22, 89, 120]).

The next major step in infinite dimensional topology, as it turned out later, was made within finite dimensional topology. In his fundamental thesis [20] M. Bestvina characterized the (finite dimensional!) universal Menger compactum μ^n and developed the theory of μ^n-manifolds. It is crucial to emphasize that the characterization in this case was given within the theory of ANE(n)-spaces in terms of the n-dimensional version of the above mentioned strong universality property.

An important philosophical consequence of Toruńczyk's and Bestvina's works can roughly be expressed as follows: the n-dimensional universal Menger compactum μ^n and not the standard n-dimensional cube I^n should be considered as a true n-dimensional counterpart of the Hilbert cube I^{ω}; these two spaces cannot be distinguished by means of "n-dimensional tests". There even is a profound analogy between the I^{ω}-manifold and μ^n-manifold theories. Almost every statement from one of these theories has its counterpart in another. The same applies to the theories of shape and n-shape. Several examples of such pairs of statements can be found in Sections 3 and 4. These results led, in turn, to a development of the theory of Nöbeling manifolds (shortly, ν^n-manifolds) [51] as "finite dimensional" analogs of R^{ω}-manifolds.

Is it really necessary to present all particular results of each of the mentioned theories separately in order to have a nonzero probability that a nonexperienced reader, specifically a beginner in the field, understands basic ideas and tastes a flavor of typical techniques in the subject? Is there a general point of view incorporating both infinite and finite dimensional theories?

Below the reader will find an attempt at such an approach. We present basics of an "(in)finite dimensional topology modulo a CW complex", define major concepts and provide necessary constructions. Some of the statements are accompanied by proofs (or their sketches). For a given complex L (actually for a class $[L]$ of complexes associated with L) we develop a theory of ANE($[L]$)-spaces and introduce the corresponding homotopy machinery based on the relation of $[L]$-homotopy. Following the classical guidelines (known as "Edwards' strategy") we discuss the problem of existence of model $[L]$-dimensional strongly $[L]$-universal ANE($[L]$)-spaces. Generally speaking, for each L, one expects to have two model spaces: $\mu^{[L]}$ – the Menger compactum modulo L and $\nu^{[L]}$ – the Nöbeling space modulo L. For which L do these objects exist? It turns out that noncompact model spaces, the Nöbeling spaces modulo L, can be constructed for each finitely dominated complex. Whereas the Menger compacta modulo L are constructed only for finite L. Such objects do not always exist as shown by the case $L = K(\mathbb{Z}, n)$. An important problem whether compacta $\mu^{[L]}$ can be constructed for finitely dominated L also remains open. Two particular complexes, namely the point $\{\mathrm{pt}\}$ and the n-dimensional sphere S^n, produce the theories of ANE- and ANE(n)-spaces, respectively. We will see that the $[\{\mathrm{pt}\}]$-homotopy coincides with the usual homotopy, whereas the $[S^n]$-homotopy is precisely the n-homotopy in the sense of J.H.C. Whitehead [151] (see also [95]). The universal model spaces in these two particular cases produce the basic objects mentioned above: $I^\omega = \mu^{[\mathrm{pt}]}$, $R^\omega = \nu^{[\mathrm{pt}]}$, $\mu^n = \mu^{[S^n]}$ and $\nu^n = \nu^{[S^n]}$. After such a general discussion we return to a traditional form of presentation of classical topics of "genuine" infinite dimensional topology and shape theory. A small hope remains that a so organized overview of the subject will be helpful at least to some of the readers.

Regrettably several important topics (such as infinite dimensional spaces arising in classical dimension theory, geometry of Banach spaces, classical n-manifolds, algebraic aspects of shape theory, applications of K-theory etc.) are left without discussion.

I would like to express my sincere thanks to the editors of this handbook. I am especially grateful to Dick Sher, who has carefully read the entire manuscript and has made many valuable comments and suggestions.

2. "Infinite dimensional topology" modulo a complex

All spaces, with the exception of Section 5, are *metrizable and separable*. Maps are continuous. We call separable completely metrizable spaces the *Polish spaces*. All complexes considered in this paper are (unless it is explicitly indicated otherwise) *countable and locally finite CW complexes*. By $\vee$ we denote the wedge operation, which unless otherwise specified is taken with arbitrary base points.

2.1. *Extension types of complexes*

For spaces X and L, the notation $L \in \mathrm{A(N)E}(X)$ means that every map $f : A \to L$, defined on a closed subspace A of X, admits an extension $\tilde{f} : X \to L$ (respectively, $\tilde{f} : G \to L$) over X (respectively, over a neighborhood G of A in X).

Next we introduce a relation $\leqslant$ for complexes. Following [76], we say that $L \leqslant K$ if for each space X the condition $L \in \mathrm{AE}(X)$ implies the condition $K \in \mathrm{AE}(X)$. Equivalence classes of complexes with respect to this relation are called *extension types*. The above defined relation $\leqslant$ creates a partial order in the class of extension types. This partial order will still be denoted by $\leqslant$ and extension type with representative L will be denoted by $[L]$. Note that under these definitions the class of all extension types has both maximal and minimal elements. The minimal element is the extension type of the 0-dimensional sphere S^0 (i.e., the two-point discrete space) and the maximal element is obviously the extension type of the one-point space {pt} (or, equivalently, of any contractible complex).

It is important to understand the above introduced partial order on the set of all extension types of complexes. We present some examples.

EXAMPLE 2.1.

(i) Clearly, $S^n \in \mathrm{AE}(X) \Leftrightarrow \dim X \leqslant n$.

(ii) Similarly, $K(G,n) \in \mathrm{AE}(X) \Leftrightarrow \dim_G X \leqslant n$. Here $\dim_G X$ stands for the *cohomological dimension*[1] of X with coefficients in an abelian group G and $K(G,n)$ denotes the corresponding *Eilenberg–MacLane complex*, i.e., a complex satisfying the following conditions: $\pi_n(K(G,n)) = G$ and $\pi_k(K(G,n)) = 0$ for each $k \neq n$.

(iii) Obviously $[S^n] \leqslant [K(\mathbb{Z},n)]$, but $[S^n] \neq [K(\mathbb{Z},n)]$. The last part follows from [74] (for $n=3$) and [87] (for $n=2$).

(iv) Clearly

$$[S^0] < [S^1] < \cdots < [S^n] < [S^{n+1}] < \cdots$$

and

$$[\mathrm{pt}] = \sup\{[S^n]:\ n = 1,2,\ldots\}.$$

(v) Similarly

$$[K(\mathbb{Z},1)] < [K(\mathbb{Z},2)] < \cdots < [K(\mathbb{Z},n)] < [K(\mathbb{Z},n+1)] < \cdots$$

and

$$[\mathrm{pt}] = \sup\{[K(\mathbb{Z},n)]:\ n = 1,2,\ldots\}.$$

(vi) $\min\{[L],[K]\} = [L \vee K]$.

(vii) $[S^0] = [L] \Leftrightarrow L$ is not connected.

(viii) $[S^1] \leqslant [L] \Leftrightarrow L$ is connected.

EXAMPLE 2.2.

(i) $[L] \leqslant [\Sigma(L)]$. Moreover, if $L \in \mathrm{AE}(X)$, then $\Sigma(L) \in \mathrm{AE}(X \times [0,1])$ (see [75]).

(ii) $L \in \mathrm{AE}(X) \Rightarrow \Sigma(L) \in \mathrm{AE}(\Sigma(X))$.

EXAMPLE 2.3. Since $[S^0] \leqslant [L]$ for any complex L, it follows from Example 2.2 that[2]

$$[\mathrm{pt}] = \sup\{[\Sigma^n(L)]:\ n = 1,2,\ldots\}.$$

Compare with Example 2.1(iv).

[1] See [83] for a comprehensive discussion of the theory of cohomological dimension.

[2] $\Sigma^n(L)$ denotes the iterated suspension of L.

EXAMPLE 2.4.

(i) It follows from Example 2.1(vii), (viii) that there is no complex L such that $[S^0] < [L] < [S^1]$.

(ii) There exists a 2-dimensional compactum X such that $\dim_G X = 1$ for some group G. This implies that $[S^1] < [S^2 \vee K(G,1)] < [S^2]$.

(iii) Let $L = M(\mathbb{Z}_2, n+1) \vee S^{n+1}$, where $M(\mathbb{Z}_2, n+1)$ is the Moore complex. Obviously, $[S^n] < [L] < [S^{n+1}]$.

(iv) If $[L] \leqslant [S^n]$, then $[L] = [L^{(n)} \vee S^n]$.

EXAMPLE 2.5. It follows from [79] that the extension type $[\mathbb{R}P^2]$ of the projective plane is not comparable with $[S^n]$ for any $n \geqslant 2$.

The Homotopy Extension Theorem implies the following trivial observation.

PROPOSITION 2.1. *If L and K are homotopy equivalent complexes, then* $[L] = [K]$.

Observe that $[S^n \vee S^{n+1}] = [S^n]$. This shows that homotopy inequivalent complexes might have the same extension type.

The following result shows that if only compact spaces are used in the definition of the relation $\leqslant$ (see the beginning of Section 2.1), then every (not necessarily countable) complex is equivalent to a wedge of countable complexes.

PROPOSITION 2.2 ([77]). *Let L be a complex, then*

$$[L] = \Big[\bigvee\{K\colon\ K \text{ is a countable complex such that } L \leqslant K\}\Big].$$

2.2. *Extension dimension*

The next step of our program is to define the *extension dimension*, $\mathrm{ed}(X)$, of a space X.

DEFINITION 2.1 ([53,76]). The extension dimension of a space X is the minimum of extension types of complexes L satisfying the relation $L \in \mathrm{AE}(X)$:

$$\mathrm{ed}(X) = \min\{[L]\colon\ L \in \mathrm{AE}(X)\}.$$

Of course extension dimension can be defined for larger classes of spaces. It was shown in [53] that using a spectral technique [51] all general questions can be reduced to similar ones for Polish spaces.

PROPOSITION 2.3 ([53]). *For a given complex L there is a Polish space $X^{[L]}$ such that* $\mathrm{ed}(X) = [L]$. *If, in addition, L is finitely dominated, we may assume that $X^{[L]}$ is compact.*

Generally speaking the extension dimension of a Polish space is not always defined (it is, if we consider arbitrary, not necessarily countable, complexes [76]).

The following question[3] remains open.

PROBLEM 2.1. Is the extension dimension of a metrizable compactum (or, more generally, of a Polish space) generated by a countable complex?

Several standard statements of the classical dimension theory remain true for the extension dimension.

PROPOSITION 2.4 ([53, 76, 145]). *If* $Y \subseteq X$, *then* $\mathrm{ed}(Y) \leqslant \mathrm{ed}(X)$.

PROPOSITION 2.5 ([53, 76, 145]). *If* $X = \bigcup\{X_k\colon k \in \omega\}$, *where* X_k *is a closed subspace of* X *such that* $\mathrm{ed}(X_k) \leqslant [L]$, $k \in \omega$, *then* $\mathrm{ed}(X) \leqslant [L]$.

PROPOSITION 2.6 ([53, 123]). *Let* $X = \lim \mathcal{S}$, *where* $\mathcal{S} = \{X_k, p_k^{k+1}, \omega\}$ *be an inverse sequence consisting of Polish spaces such that* $\mathrm{ed}(X_k) \leqslant [L]$ *for each* $k \in \omega$. *Then* $\mathrm{ed}(X) \leqslant [L]$.

PROPOSITION 2.7. *Let* K *be a locally compact polyhedron.*

(a) *If* $\dim K = n$, *then* $\mathrm{ed}(K) = [S^n]$.

(b) *If* K *is infinite dimensional, then* $\mathrm{ed}(K) = [\mathrm{pt}]$.

PROOF. First we show that $\mathrm{ed}(I^n) = [S^n]$. Clearly $\mathrm{ed}(I^n) \leqslant [S^n]$. Let L be a complex such that $L \in \mathrm{AE}(I^n)$. By Propositions 2.4 and 2.5, $L \in \mathrm{AE}(K)$ for each at most n-dimensional compact polyhedron. Further, if Z is an at most n-dimensional metrizable compactum, then, by Freudental's theorem, Z is the limit space of an inverse sequence $\{K_i, p_i^{i+1}\}$ consisting of at most n-dimensional compact polyhedra. By the above remark and by Proposition 2.6, $L \in \mathrm{AE}(Z)$. In particular, $L \in \mathrm{AE}(\mu^n)$, where μ^n denotes the n-dimensional universal Menger compactum. Now, if Y is an at most n-dimensional space, we may assume that $Y \subseteq \mu^n$. Consequently, by Proposition 2.4, $L \in \mathrm{AE}(Y)$. In other words $S^n \in \mathrm{AE}(X) \Rightarrow L \in \mathrm{AE}(X)$ for each space X. According to the definition of the above partial order, this means precisely that $[S^n] \leqslant [L]$ for each L satisfying the condition $L \in \mathrm{AE}(I^n)$. Therefore $\mathrm{ed}(I^n) = [S^n]$. This immediately implies that $\mathrm{ed}(K) = [S^n]$ for each n-dimensional locally compact polyhedron.

Now assume that K is an infinite dimensional locally compact polyhedron and L is a complex such that $L \in \mathrm{AE}(K)$. Since every contractible complex generates the extension type of the one-point space $\{\mathrm{pt}\}$, it suffices to prove that L is contractible. Assuming that this is not the case, L must have at least one nontrivial homotopy group, say $\pi_n(L) \neq 0$. Let $f : S^n \to L$ be a map generating a nonzero element of $\pi_n(L)$. Clearly K contains the $(n+1)$-dimensional disk B^{n+1} and since $L \in \mathrm{AE}(K)$ we conclude that the map $f : \partial B^{n+1} = S^n \to L$ has an extension over B^{n+1}. This contradicts the nontriviality of the homotopy class $[f] \in \pi_n(L)$ and finishes the proof. □

In the light of the previous statement the following result appears to be somewhat unexpected.

[3] Compare with Proposition 2.2

PROPOSITION 2.8 ([56]). *For an arbitrary complex L for which there exists a metrizable compactum X such that* $\mathrm{ed}(X) = [L]$, *assuming Jensen's principle, there exists a differentiable, countably compact, perfectly normal and hereditarily separable* 4*-manifold* M^L *such that* $\mathrm{ed}(M^L) = [L]$.

2.3. *Absolute extensors modulo a complex*

In this subsection we introduce the notion of absolute extensors modulo L.

DEFINITION 2.2. We say that a space X is an absolute (neighborhood) extensor modulo L, or shortly that X is an A(N)E($[L]$)-space (notation: $X \in \mathrm{A(N)E}([L])$) if $X \in \mathrm{A(N)E}(Y)$ for each space Y with $\mathrm{ed}(Y) \leqslant [L]$.

REMARK 2.1. It follows from the above definitions that $L \in \mathrm{AE}([L])$.

Example 2.1(vii), (viii) explains the following statement.

PROPOSITION 2.9. *If L is not connected, then the class of* A(N)E($[L]$)*-spaces coincides with the class of* A(N)E($[S^0]$)*-spaces. If L is connected, then every* A(N)E($[L]$)*-space is an* A(N)E($[S^1]$)*-space.*

PROPOSITION 2.10. *If* $[L] \leqslant [P]$, *then the class of* A(N)E($[P]$)*-spaces is contained in the class of* A(N)E($[L]$)*-spaces.*

PROPOSITION 2.11. *An open subspace of an* ANE($[L]$)*-space is an* ANE($[L]$)*-space.*

PROPOSITION 2.12. *A retract of an* A(N)E($[L]$)*-space is an* A(N)E($[L]$)*-space.*

EXAMPLE 2.6. Observe that:
(i) ANE($[\{\mathrm{pt}\}]$)-spaces are precisely ANE-spaces.
(ii) ANE($[S^n]$)-spaces are precisely ANE(n)-spaces, i.e., LC^{n-1}-spaces.

Next we introduce the concepts of $[L]$-soft and $[L]$-invertible maps [53].

DEFINITION 2.3. A map $f : X \to Y$ is said to be $[L]$-*soft*, if for each space B with $\mathrm{ed}(B) \leqslant [L]$, for each closed subset A of it, and for any two maps $g : A \to X$ and $h : B \to Y$ such that $fg = h/A$, there is a map $k : B \to X$ satisfying the conditions $k/A = g$ and $fk = h$. If the above condition is satisfied in the cases when $B = \emptyset$ we say that f is $[L]$-*invertible*.

Clearly, every $[L]$-soft map is $[L]$-invertible. $[L]$-invertible maps are surjective, while $[L]$-soft maps are open [53, Remark 5.10]. We list the following immediate corollaries.

PROPOSITION 2.13. *The following conditions are equivalent for an $[L]$-soft map* $f : X \to Y$:
(i) $X \in \mathrm{A(N)E}([L])$.
(ii) $Y \in \mathrm{A(N)E}([L])$.

In particular, $X \in \mathrm{AE}([L])$ if and only if the constant map $X \to \{\mathrm{pt}\}$ is $[L]$-soft.

The following statement is important for our considerations.

THEOREM 2.1 (Proposition 5.9 [53]). *Let L be a complex and X be a (Polish) space. Then there exists an $[L]$-soft map $f_X : Y_X \to X$, where Y_X is a (Polish) space with $\mathrm{ed}(Y_X) = [L]$. If, in addition, A is a closed subset of X such that $\mathrm{ed}(A) \leqslant [L]$, then we may assume that $f_X / f_X^{-1}(A) : f_X^{-1}(A) \to A$ is a homeomorphism.*

COROLLARY 2.1 ([53, 123]). *For a given complex L there exists a Polish $\mathrm{AE}([L])$-space $X^{[L]}$ such that $\mathrm{ed}(X^{[L]}) = [L]$ and $X^{[L]}$ contains a (closed) topological copy of any (Polish) space Y with $\mathrm{ed}(Y) \leqslant [L]$.*

PROOF. Let $f : X^{[L]} \to I^{\omega}$ be an $[L]$-soft map constructed in Theorem 2.1. Since I^{ω} is an $\mathrm{AE}([\{\mathrm{pt}\}])$-space, by Proposition 2.10, I^{ω} is an $\mathrm{AE}([L])$-space. Then, by Proposition 2.13, $X^{[L]}$ is an $\mathrm{AE}([L])$-space. Now let Y be a space such that $\mathrm{ed}(Y) \leqslant [L]$. We may assume that $Y \subseteq I^{\omega}$. Since f is $[L]$-soft the inclusion map $Y \hookrightarrow I^{\omega}$ admits a lifting $i_Y : Y \to X^{[L]}$. Clearly i_Y is an embedding. □

COROLLARY 2.2 ([53, 122]). *Every space X has a completion $\widetilde{X}$ such that $\mathrm{ed}(\widetilde{X}) = \mathrm{ed}(X)$.*

There are spaces without cohomological dimension preserving compactifications [86]. This implies the following.

PROPOSITION 2.14. *There is a space X such that $\mathrm{ed}(X) < \mathrm{ed}(\widetilde{X})$ for every compactification $\widetilde{X}$ of X.*

Consequently we cannot assume in Corollary 2.1 that the space $X^{[L]}$ is compact. This still leaves the following question open.

PROBLEM 2.2. Does there exist a compact space $X^{[L]}$ such that $\mathrm{ed}(X^{[L]}) = [L]$ and which contains topological copy of every compact space Y with $\mathrm{ed}(Y) \leqslant [L]$?

For finitely dominated complexes, the situation is simpler. A helpful observation here is the following statement.

THEOREM 2.2 ([55]). *Let L be a finitely dominated connected locally compact complex. Then the following conditions are equivalent for any (not necessarily normal[4]) space X:*

(i) $L \in \mathrm{AE}(X)$.

(ii) $L \in \mathrm{AE}(\beta X)$, *where βX denotes the Stone–Čech compactification of X.*

Although Theorem 2.1 allows us to establish several interesting facts, the lack of properness of the $[L]$-soft map constructed in this theorem does not produce extension dimension

[4] See Definition 5.1.

preserving compactifications or compact universal spaces of a given extension dimension. Our next statement provides a *proper* $[L]$-invertible (as it follows from Proposition 2.16 there are no proper $[L]$-soft maps of $[L]$-dimensional spaces) map.

THEOREM 2.3. *Let L be a finitely dominated complex and X be a (Polish) space. Then there exists a proper $[L]$-invertible map $f_X : Y_X \to X$, where Y_X is a (Polish) space such that* $\mathrm{ed}(Y_X) = [L]$. *If, in addition, A is a closed subset of X such that* $\mathrm{ed}(A) \leqslant [L]$, *then we may assume that $f_X / f_X^{-1}(A) : f_X^{-1}(A) \to A$ is a homeomorphism.*

PROOF. Let $\mathcal{A}$ denote the set of all maps $\{f_t\colon t \in T\}$ such that domain $\mathrm{Dom}(f_t)$ is a Polish space, $\mathrm{ed}(\mathrm{Dom}(f_t)) \leqslant [L]$ and $\mathrm{Ran}(f_t) \subseteq I^\omega$. Let $Y = \bigoplus\{\mathrm{Dom}(f_t)\colon t \in T\}$. Clearly, $\mathrm{ed}(Y) \leqslant [L]$. Consider the map $f : Y \to I^\omega$, which coincides with the map f_t on $\mathrm{Dom}(f_t)$ for each $t \in T$, and its extension $\beta f : \beta Y \to I^\omega$. Since L is finitely dominated, by Theorem 2.2, we have $\mathrm{ed}(\beta Y) = \mathrm{ed}(Y) \leqslant [L]$. Consequently, by [53, Theorem 4.4], βY is the limit space of a Polish spectrum $\mathcal{S}_Y = \{Y_\alpha, q_\alpha^\beta, A\}$ consisting of compact metrizable spaces Y_α such that $\mathrm{ed}(Y_\alpha) \leqslant [L]$ for each $\alpha \in A$. Since I^ω is a metrizable compactum there exist [51] an index $\alpha \in A$ and a map $f_\alpha : Y_\alpha \to I^\omega$ such that $\beta f = f_\alpha q_\alpha$.

We claim that the map $f_\alpha \to I^\omega$ is $[L]$-invertible. Indeed, let $g : Z \to I^\omega$ be a map such that $\mathrm{ed}(Z) \leqslant [L]$. By the definition of $\mathcal{A}$, there is an index $t \in T$ such that $f_t = g$. Let $i_t : \mathrm{Dom}(f_t) = Z \to Y$ denote the corresponding embedding. Clearly $f_t = f i_t$. Consequently, the composition $h = q_\alpha i_t : Z \to Y_\alpha$ lifts the map g, i.e., $f_\alpha h = g$. This shows that f_α is $[L]$-invertible.

If X is an arbitrary (Polish) space, then embed X into I^ω and consider the $[L]$-invertible map $f_\alpha : Y_\alpha \to I^\omega$ constructed above. Clearly, $\mathrm{ed}(Y_X) \leqslant \mathrm{ed}(Y_\alpha) \leqslant [L]$, where $Y_X = f_\alpha^{-1}(X)$. It only remains to note that the map $f_X = f_\alpha / Y_X : Y_X \to X$ is proper and L-invertible and $\mathrm{ed}(Y_X) = [L]$. □

PROPOSITION 2.15. *Let L be a finitely dominated complex,* $\mathrm{ed}(X) \leqslant [L]$ *and $f : X \to Y$ be a map of a Polish space X into a metrizable compactum Y. Then there is a map $\tilde{f} : \tilde{X} \to Y$ such that $\tilde{X}$ is a metrizable compactification of X with* $\mathrm{ed}(\tilde{X}) \leqslant [L]$ *and $f = \tilde{f}/X$.*

COROLLARY 2.3. *Let L be a finitely dominated complex. Then there exists a metrizable compactum $X^{[L]}$ such that* $\mathrm{ed}(X_{[L]}) = [L]$ *and the following conditions are equivalent for any space X:*

(i) $\mathrm{ed}(X) \leqslant [L]$.

(ii) *X admits an embedding into $X^{[L]}$.*

PROOF. Let $f : X^{[L]} \to I^\omega$ be an $[L]$-invertible map constructed in Theorem 2.3. We claim that $X^{[L]}$ is the desired compactum. First we show that $\mathrm{ed}(X^{[L]}) = [L]$. Assuming that this is not true, and having in mind that $\mathrm{ed}(X^{[L]}) \leqslant [L]$, we conclude that there is a complex K such that $\mathrm{ed}(X^{[L]}) \leqslant [K] < [L]$. This, by the definition of the order on extension types, means that there is a Polish space Y such that $L \in \mathrm{AE}(Y)$ but $K \notin \mathrm{AE}(Y)$. Embed Y into I^ω. Since $\mathrm{ed}(Y) \leqslant [L]$ and since f is $[L]$-invertible, there is a map $g : Y \to X^{[L]}$ such that

$fg = \mathrm{id}_Y$. Obviously, g is an embedding and hence $\mathrm{ed}(Y) = \mathrm{ed}(g(Y)) \leqslant \mathrm{ed}(X^{[L]}) \leqslant [K]$, which is impossible by the choice of Y ($K \notin \mathrm{AE}(Y)$). Thus, $\mathrm{ed}(X^{[L]}) = [L]$.

(i) $\Rightarrow$ (ii). We may assume that $X \subset I^\omega$. Since $\mathrm{ed}(X) \leqslant [L]$ and since f is L-invertible, there is a map $g : X \to X^P$ such that $fg = \mathrm{id}_X$. Clearly, g is an embedding.

The implication (ii) $\Rightarrow$ (i) is trivial. □

The following corollary does not contradict Proposition 2.14 and follows from Theorem 2.2.

COROLLARY 2.4 (I.A. Shvedov, see [82, 111]). *Let L be a finitely dominated complex and* $\mathrm{ed}(X) \leqslant [L]$. *Then there exists a compactification $\widetilde{X}$ of X such that* $\mathrm{ed}(\widetilde{X}) \leqslant [L]$.

PROPOSITION 2.16. *There is no $[L]$-soft map $f : X \to Y$ where X is a compactum with* $\mathrm{ed}(X) \leqslant [L]$ *and Y is a connected* ANE-*compactum with* $\mathrm{ed}(Y) > [L]$.

PROOF. First consider the case of nonconnected L. By Example 2.1(vii), $[L] = [S^0]$. Then $\dim X = 0$ and, by [53, Remark 5.10] and [51, Corollary 6.1.27], f is open. Therefore Y must also be zero-dimensional and this contradicts our assumptions.

Suppose now that L is a connected complex. By Example 2.1(viii) and Proposition 2.13, the fibers of f are connected. This in turn implies that X is connected. By the following proposition, f is a cell-like map. By [80], f is approximately invertible and hence $\mathrm{ed}(Y) \leqslant \mathrm{ed}(X) \leqslant [L]$. Contradiction. □

PROPOSITION 2.17 ([61]). *Every non-constant $[L]$-soft map $f : X \to Y$, where X is a connected compactum such that* $\mathrm{ed}(X) \leqslant [L]$, *is cell-like.*

2.4. *Z-sets*

2.4.1. *Limitation topology on spaces of maps.* The collection of all open (countable) covers of a space X is denoted by $\mathrm{cov}(X)$. Two maps $f, g : Y \to X$ are called *$\mathcal{U}$-close* (notation: $(f, g) < \mathcal{U}$), where $\mathcal{U} \in \mathrm{cov}(X)$, if for each point $y \in Y$ there exists an element $U \in \mathcal{U}$ such that $f(y), g(y) \in U$. Denote by $B(f, \mathcal{U})$ the collection of all $\mathcal{U}$-close to f maps. The *limitation topology* on the set $C(Y, X)$ of all continuous maps of Y into X is now defined as follows. A set $G \subseteq C(Y, X)$ is open if for each $f \in G$ there is an open cover $\mathcal{U} \in \mathrm{cov}(\mathcal{X})$ such that $B(f, \mathcal{U}) \subseteq G$. Note that the sets $B(f, \mathcal{U})$ are not necessarily open.

We say that a map $f : Y \to X$ is a *$\mathcal{U}$-map*, where $\mathcal{U} \in \mathrm{cov}(Y)$, if there exists an open cover $\mathcal{V} \in \mathrm{cov}(X)$ such that the cover $f^{-1}(\mathcal{V}) = \{f^{-1}(V)\colon\ V \in \mathcal{V}\}$ refines $\mathcal{U}$.

LEMMA 2.1. *Let $\mathcal{U} \in \mathrm{cov}(Y)$. Then the set of all $\mathcal{U}$-maps is open in $C(Y, X)$.*

LEMMA 2.2. *Let ρ be a complete metric on a Polish space Y and suppose a sequence $\{\mathcal{U}_n\colon\ n \in N\} \subseteq \mathrm{cov}(Y)$ is chosen so that $\mathrm{diam}_\rho U < \frac{1}{n}$ for each $U \in \mathcal{U}_n$ and each $n \in N$. Then a map $f : Y \to X$ is a closed embedding* (*i.e., an embedding with closed image*) *if* *and only if f is a $\mathcal{U}_n$-map for each $n \in N$. In particular, the set of all closed embeddings of Y into X is a G_δ-subset of $C(Y, X)$.*

Every bounded metric ρ on a space X generates the sup-metric $\check{\rho}$ on $C(Y, X)$:

$$\check{\rho}(f, g) = \sup\{\rho(f(y), g(y)) \colon y \in Y\}, \quad f, g \in C(Y, X).$$

The topology generated by the metric $\check{\rho}$ on the set $C(Y, X)$ is called the *topology of ρ-uniform convergence.*

Let $\mathrm{Metr}(X)$ denote the collection of all bounded metrics on a space X compatible with the topology of X. The collection $\{\check{\rho}\colon \rho \in \mathrm{Metr}(X)\}$ generates the *topology of uniform convergence with respect to all bounded metrics* on $C(Y, X)$. Obviously, this topology is stronger than the topology of ρ-uniform convergence for each $\rho \in \mathrm{Metr}(X)$.

LEMMA 2.3. *The limitation topology coincides with the topology of uniform convergence with respect to all bounded metrics.*

The following statement provides another description of the limitation topology.

LEMMA 2.4. *Let $\rho \in \mathrm{Metr}(X)$ and $f \in C(Y, X)$. The collection $\{B_\rho(f, \alpha)\colon \alpha \in C(X, (0, \infty))\}$, where $B_\rho(f, \alpha) = \{g \in C(Y, X)\colon \check{\rho}(f(y), g(y)) \leqslant \alpha(f(y))$ for each $y \in Y\}$, forms a local basis at f in $C(Y, X)$.*

The following statement expresses a very important and frequently used property of the limitation topology.

PROPOSITION 2.18. *Let X be a Polish space, F be a subspace of the space $C(Y, X)$ and the set G_n be open in $C(Y, X)$ for each $n \in N$. If the intersection $G_n \cap F$ is dense in F for each $n \in N$, then $F \subseteq \mathrm{cl}_{C(Y,X)}((\bigcap G_n) \cap F_\rho)$, where F_ρ denotes the closure of F in the topology of ρ-uniform convergence and ρ is any bounded metric on X. In particular, $C(Y, X)$ has the Baire property.*

Proposition 2.18 can be used to characterize maps between Polish spaces which are approximable by homeomorphisms. Such maps are called near-homeomorphisms. More formally, a map $f\colon Y \to X$ is a *near-homeomorphism* if, given $\mathcal{U} \in \mathrm{cov}(X)$, there is a homeomorphism $h_{\mathcal{U}}\colon Y \to X$ which is $\mathcal{U}$-close to f.

THEOREM 2.4 (Bing's Shrinking criterion). *A map $f\colon Y \to X$ between Polish spaces is a near-homeomorphism if and only if $f(Y)$ is dense in X and the following condition is satisfied:*

($\star$) *For each $\mathcal{U} \in \mathrm{cov}(Y)$ and $\mathcal{V} \in \mathrm{cov}(X)$ there exist an open cover $\mathcal{W} \in \mathrm{cov}(X)$ and a homeomorphism $h\colon Y \to Y$ such that $fh \in B(f, \mathcal{V})$ and $hf^{-1}(\mathcal{W}) \prec \mathcal{U}$.*

COROLLARY 2.5. *A closed surjection $f\colon Y \to X$ between Polish spaces is a near-homeomorphism if and only if the following condition is satisfied:*

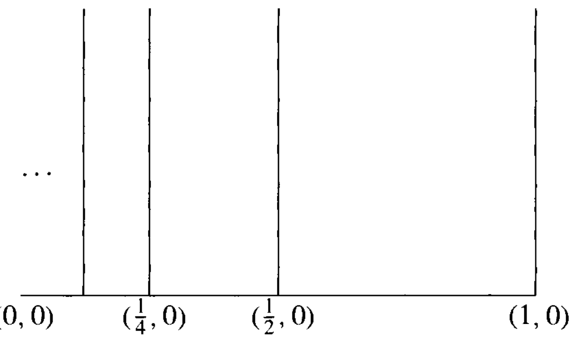
(⋆⋆) *For each* $\mathcal{U} \in \text{cov}(Y)$ *and each* $\mathcal{V} \in \text{cov}(X)$ *there exists a homeomorphism* $h : Y \to Y$ *such that* $fh \in B(f, \mathcal{V})$ *and the collection* $\{hf^{-1}(x)\colon x \in X\}$ *refines* $\mathcal{U}$.

2.4.2. *Definition of (strong) Z-sets.* As usual a closed subset A of a space X is said to be a *Z-set* in X if the set $\{f \in C(X, X)\colon f(X) \cap A = \emptyset\}$ is dense in the space $C(X, X)$. If the set $\{f \in C(X, X)\colon \text{cl}_X f(X) \cap A = \emptyset\}$ is dense in $C(X, X)$, then we say that A is a *strong Z-set* in X.

The concepts of the Z-set and the strong Z-set differ even for very simple spaces [22]. Consider the subset

$$X = ([0,1] \times \{0\}) \bigcup \left(\bigcup_{n=1}^{\infty} \left\{ \left\{ \frac{1}{n} \right\} \times [0,1] \right\} \right)$$

of the plane:



It is not hard to see that X is a AE-space. Moreover, X is even locally compact at each point other than the origin $(0, 0)$. For each $t \in [0, 1)$ consider a map $h_t : X \to X$ determined by the following properties:
- $h_t(1, 0) = (1, 0)$.
- $h_t(\frac{1}{n}, 1) = (\frac{1}{n}, 1)$.
- h_t linearly shrinks the segment $[0, 1] \times \{0\}$ onto the segment $[t, 1] \times \{0\}$.
- h_t linearly expands the segment $\{\frac{1}{n}\} \times [t, 1]$ onto the segment $\{\frac{1}{n}\} \times [0, 1]$.
- h_t linearly maps the segment $\{\frac{1}{n}\} \times [0, t]$ onto the segment $[\frac{1}{n}, t + \frac{1-t}{n}] \times \{0\}$ (sending $(\frac{1}{n}, t)$ onto $(\frac{1}{n}, 0)$ and $(\frac{1}{n}, 0)$ onto $(t + \frac{1-t}{n}, 0)$).

Clearly, the identity map id_X is in the closure (in the space $C(X, X)$) of the set $\{h_t\colon t \in (0, 1)\}$, and since $h_t(X) \subseteq X - \{(0, 0)\}$ it follows that the set $\{(0, 0)\}$ is a Z-set in X. At the same time $\{(0, 0)\}$ is not a strong Z-set in X. To see this, observe that $X - G$ is disconnected for sufficiently small neighborhoods G of the point $(0, 0)$.

LEMMA 2.5. *Let L be a finitely dominated complex and X be an* ANE$([L])$*-space such that* $\text{ed}(A) \leqslant [L]$. *For each open cover* $\mathcal{U} \in \text{cov}(X)$ *there exist a locally compact space* $K_{\mathcal{U}}$ *with* $\text{ed}(K_{\mathcal{U}}) \leqslant [L]$ *and two maps* $\alpha : X \to K_{\mathcal{U}}$ *and* $\beta : K_{\mathcal{U}} \to X$ *such that the composition* $\beta\alpha$ *is* $\mathcal{U}$*-close to* id_X.

PROOF. Assume that X is closed in R^ω and consider a proper $[L]$-invertible map $p: Y \to R^\omega$ where Y is a Polish space with $\mathrm{ed}(Y) \leqslant [L]$. Also let $q: p^{-1}(V) \to X$ be an extension of the map $p/p^{-1}(X)$ and $\mathcal{U}' = \{U' \colon U \in \mathcal{U}\}$ be a collection of open subsets of V such that

$$U' \cap X = U \quad \text{and} \quad p^{-1}(U') \subseteq q^{-1}(U) \quad \text{for each } U \in \mathcal{U}.$$

Since $S = \bigcup\{U' \colon U \in \mathcal{U}\}$ is a Polish ANE-space there exist a locally compact polyhedron K and two maps $\gamma: S \to K$ and $\delta: K \to S$ such that $\delta\gamma$ is $\mathcal{U}'$-close to id_S. By Theorem 2.3, there is a proper $[L]$-invertible map $h: K_{\mathcal{U}} \to K$ of a (locally compact) space $K_{\mathcal{U}}$ such that $\mathrm{ed}(K_{\mathcal{U}}) \leqslant [L]$. The $[L]$-invertibility of h coupled with the fact that $e(X) \leqslant [L]$ guarantees that there is a map $\alpha: X \to K_{\mathcal{U}}$ such that $h\alpha = \gamma/X$. The $[L]$-invertibility of p coupled with the fact that $e(K_{\mathcal{U}}) \leqslant [L]$ implies the existence of a map $\beta': K_{\mathcal{U}} \to p^{-1}(L)$ such that $p\beta' = \delta$. Let $\beta = q\beta'$. It only remains to observe that $\alpha\beta$ is indeed $\mathcal{U}$-close to id_X. □

PROPOSITION 2.19. *Let L be a finitely dominated complex, $X^{[L]}$ be a universal compactum in the class of spaces whose extension dimension does not exceed $[L]$ and $\{X_k^{[L]} \colon k \in \omega\}$ be a countable closed base of $X^{[L]}$ containing intersections of its finite subcollections. Then the following conditions are equivalent for a closed subset A of a Polish* ANE$([L])$*-space X:*

(i) *The set $\{f \in C(X_k^{[L]}, X) \colon f(X_k^{[L]}) \cap A = \emptyset\}$ is dense in the space $C(X_k^{[L]}, X)$ for each $k \in \omega$.*
(ii) *The set $\{f \in C(Y, X) \colon f(Y) \cap A = \emptyset\}$ is dense in the space $C(Y, X)$ for each compact space Y with* $\mathrm{ed}(Y) \leqslant [L]$.
(iii) *The set $\{f \in C(Y, X) \colon f(Y) \cap A = \emptyset\}$ is dense in the space $C(Y, X)$ for each locally compact space Y with* $\mathrm{ed}(Y) \leqslant [L]$.
(iv) *A is a Z-set in X.*

PROOF. (i) $\Rightarrow$ (ii). Let $f: Y \to X$ be a map defined on a compactum Y such that $\mathrm{ed}(Y) \leqslant [L]$. We may assume that $Y \subseteq X^{[L]}$. Since $\mathrm{ed}(X^{[L]}) = [L]$ and since X is an ANE$([L])$-space, there is a $k \in \omega$ such that $Y \subseteq X_k^{[L]}$ and f admits an extension $\tilde{f}: X_k^{[L]} \to X$. By (i), $\tilde{f}$ can be approximated by maps $\tilde{g}: X_k^{[L]} \to X$ such that $\tilde{g}(X_k^P) \cap A = \emptyset$. Then the map $g = \tilde{g}/Y: Y \to X$ is an approximation of f and $g(Y) \cap A = \emptyset$.

(ii) $\Rightarrow$ (iii). Suppose now that Y is a locally compact space such that $\mathrm{ed}(Y) \leqslant [L]$. Since Y is σ-compact, we can write $Y = \bigcup\{Y_k \colon k = 1, 2, \ldots\}$ where each Y_k is compact. Let us show that for each $k = 1, 2, \ldots$ the set

$$B_k = \{f \in C(Y, X) \colon f(Y_k) \cap A = \emptyset\}$$

is dense in the space $C(Y, X)$. Consider a map $f: Y \to X$ and an open cover $\mathcal{U} \in \mathrm{cov}(X)$. Consider an open cover $\mathcal{V} \in \mathrm{cov}(X)$ satisfying the condition of $(*)_{[L]}$ of Proposition[5] 2.27 relative to $\mathcal{U}$. By (ii), there is a map $g': Y_k \to X$ such that g' is $\mathcal{V}$-close to f/Y_k and

[5] There is no logical inconsistency here. Proof of Proposition 2.27 does not depend on Proposition 2.19.

$g'(Y_k) \cap A = \emptyset$. Consequently, by the choice of $\mathcal{V}$, the map g' has an extension $g: Y \to X$ which is $\mathcal{U}$-close to f. Clearly, $g(Y_k) \cap A = \emptyset$.

Next observe that the set B_k is open in $C(Y, X)$ and, consequently, by the Baire property of the space $C(Y, X)$ (see Proposition 2.18), the intersection

$$\{f \in C(Y, X)\colon\ f(Y) \cap A = \emptyset\} = \bigcap\{B_k\colon\ k = 1, 2, \ldots\}$$

is also dense in $C(Y, X)$.

(iii) $\Rightarrow$ (iv). Consider a map $f: Y \to X$ and an open cover $\mathcal{U} \in \operatorname{cov}(X)$. Let $\mathcal{V} \in \operatorname{cov}(X)$ be a star-refinement of $\mathcal{U}$. For an open cover $f^{-1}(\mathcal{V}) \in \operatorname{cov}(Y)$, by Lemma 2.5, there exist a locally compact space K and maps $\alpha: Y \to K$ and $\beta: K \to Y$ such that the composition $\beta\alpha$ is $f^{-1}(\mathcal{V})$-close to id_Y. By (iii), there is a map $g': K \to X$ which is $\mathcal{V}$-close to the composition $f\beta$ and such that $g'(K) \cap A = \emptyset$. Let $g = g'\alpha$. It is easy to check that g and f are $\mathcal{U}$-close and $g(X) \cap A = \emptyset$.

The implications (iv) $\Rightarrow$ (i) is trivial. □

PROPOSITION 2.20. *Let L be a finitely dominated complex, $X^{[L]}$ be a universal compactum in the class of spaces whose extension dimension does not exceed $[L]$ and $\{X_k^{[L]}\colon\ k \in \omega\}$ be a countable closed base of $X^{[L]}$ containing intersections of its finite subcollections. Then the following conditions are equivalent for a closed subset A of a Polish* ANE($[L]$)-*space X:*

(i) *The set*

$$\Big\{f \in C\Big(\bigoplus\{X_k^{[L]}\colon\ k \in \omega\}, X\Big)\colon\ \mathrm{cl}_X f\Big(\bigoplus\{X_k^{[L]}\colon\ k \in \omega\}\Big) \cap A = \emptyset\Big\}$$

is dense in the space $C(\bigoplus\{X_k^{[L]}\colon\ k \in \omega\}, X)$

(ii) *The set*

$$\Big\{f \in C\Big(\bigoplus\{Y_k\colon\ k \in \omega\}, X\Big)\colon\ \mathrm{cl}_X f\Big(\bigoplus\{Y_k\colon\ k \in \omega\}\Big) \cap A = \emptyset\Big\}$$

is dense in the space $C(\bigoplus\{Y_k\colon\ k \in \omega\}, X)$ for any compact spaces Y_k with $\mathrm{ed}(Y_k) \leqslant [L]$, $k \in \omega$.

(iii) *The set*

$$\{f \in C(Y, X)\colon\ \mathrm{cl}_X f(Y) \cap A = \emptyset\}$$

is dense in the space $C(Y, X)$ for each locally compact space Y with $\mathrm{ed}(Y) \leqslant [L]$.

(iv) *A is a strong Z-set in X.*

PROOF. In order to prove the only nontrivial implication (ii) $\Rightarrow$ (iii), represent Y as the countable union $Y = \bigcup\{Y_k\colon\ k \in \omega\}$ of compact subsets such that $Y_k \subseteq \mathrm{int}(Y_{k+1})$ for each $k \in \omega$. Let $T_1 = \bigcup\{Y_{2k} - \mathrm{int}(Y_{2k-1})\colon\ k \in \omega\} \cup Y_0$ and $T_2 = \bigcup\{Y_{2k+1} - \mathrm{int}(Y_{2k})\colon\ k \in \omega\}$. By (ii) and Proposition[6] 2.27, the sets $A_i = \{f \in C(Y, X)\colon\ \mathrm{cl}_X f(T_1) \cap A = \emptyset\}$, $i = 1, 2$, are

[6] See footnote 5.

both dense and open in $C(Y, X)$. Finally, by Proposition 2.18, their intersection $A_1 \cap A_2$ is also dense in $C(Y, X)$. □

COROLLARY 2.6. *Every Z-set in a locally compact* ANE($[L]$)*-space is a strong Z-set in it.*

2.5. *Strongly $[L]$-universal spaces*

In this subsection we discuss the main objects of "infinite dimensional topology modulo L". There are two types of such spaces: $\mu^{[L]}$ – Menger compactum modulo L, and $\nu^{[L]}$ – Nöbeling space modulo L.

2.5.1. *Locally compact case.* Below we construct Menger compacta $\mu^{[L]}$ modulo L.

DEFINITION 2.4. For a given complex L we say that a space X satisfies the *disjoint disks property modulo L* (notation: $[L]$-DDP) if the set

$$\{f \in C(Y \times \{0,1\}, X)\colon\ f(Y \times \{0\}) \cap f(Y \times \{1\}) = \emptyset\},$$

where Y is a compactum such that $\mathrm{ed}(Y) \leqslant [L]$, is dense in the space $C(Y \times \{0,1\}, X)$.

DEFINITION 2.5. For a given complex $[L]$ we say that a locally compact space X is *strongly $[L]$-universal for compact spaces* if for any locally compact space B, its closed subspace A, any open cover $\mathcal{U} \in \mathrm{cov}(X)$ and for any proper map $f : B \to X$, the restriction f/A of which is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.

PROPOSITION 2.21. *Let L be a finitely dominated complex and X be a* (*locally*) *compact* ANE($[L]$)*-space. Then the following conditions are equivalent*:

(i) *X has the* $[L]$-DDP.
(ii) *X is strongly $[L]$-universal for compact spaces.*

COROLLARY 2.7. *Every point in a locally compact* ANE($[L]$)*-space with* $[L]$-DDP *is a Z-set.*

Later we will see that the proof of Theorem 2.6, which through the inverse limit construction produces nonlocally compact spaces $\nu^{[L]}$ discussed above, cannot, without changes, be adapted to a locally compact situation. Reason – the nonexistence of proper $[L]$-soft maps raising extension dimension (Proposition 2.16). On the other hand, the $[L]$-invertibility of (proper) projections (such maps do exist – Theorem 2.3) of an inverse sequence consisting of ANE($[L]$)-spaces does not guarantee that the limit is also an ANE($[L]$)-space. The class of maps which is sufficient for our purposes is introduced in the following definition.

DEFINITION 2.6. A map $f : X \to Y$ is said to be *approximately* $[L]$-*soft*, if for each space B with $\mathrm{ed}(B) \leqslant [L]$, for each closed subset A of it, for an open cover $\mathcal{U} \in \mathrm{cov}(Y)$, and for any two maps $g : A \to X$ and $h : B \to Y$ such that $fg = h/A$, there is a map $k : B \to X$ satisfying the conditions $k/A = g$ and the composition fk is $\mathcal{U}$-close to h.

REMARK 2.2. Basic examples of approximately $[L]$-soft maps are maps between $\mathrm{ANE}([L])$-compacta whose nontrivial fibers are $\mathrm{AE}([L])$-compacta.

The standard argument [78] proves the next statement.

PROPOSITION 2.22. *Let* $\mathcal{S} = \{X_n, p_n^{n+1}, \omega\}$ *be an inverse spectrum consisting of (locally) compact* $\mathrm{ANE}([L])$*-spaces and approximately* $[L]$*-soft (proper) projections. Then the limit* $\lim \mathcal{S}$ *is an* $\mathrm{ANE}([L])$*-space and each limit projection* $p_n : \lim \mathcal{S} \to X_n$ *is approximately* $[L]$*-soft.*

The following construction [78] is based on an idea of Dranishnikov [73]. Let $\mathrm{Cone}(L)$ be the cone over a *finite* complex L and let $\lambda : L \times [0, 1] \to \mathrm{Cone}(L)$ be the quotient map. Consider the subspace $T = \{((l, t), l) \colon (l, t) \in L \times [0, 1)\}$ of $\mathrm{Cone}(L) \times L$ and an open neighborhood U of the vertex of the cone in $\mathrm{Cone}(L)$. Clearly $\widetilde{L} = (\mathrm{cl}\, U \times L) \cup T$ is a closed subspace of the product $\mathrm{Cone}(L) \times L$. Finally, let $\xi_L : \widetilde{L} \to \mathrm{Cone}(L)$ denote the restriction of the projection $\mathrm{Cone}(L) \times L \to \mathrm{Cone}(L)$ onto $\widetilde{L}$. The following statement expresses an essential property of the map ξ_L (A.N. Dranishnikov).

LEMMA 2.6. *The map* $\xi_L : \widetilde{L} \to \mathrm{Cone}(L)$ *is* $[L]$*-invertible and approximately* $[L]$*-soft.*

PROOF. The fact that ξ_L is approximately $[L]$-soft follows from Remarks 2.1 and 2.2. Let us show that ξ_L is $[L]$-invertible. Let $\varphi : B \to \mathrm{Cone}(L)$ be a map defined on a compactum B with $\mathrm{ed}(B) \leqslant [L]$. Let $\psi : \mathrm{Cone}(K) - U \to \widetilde{K}$ be the unique section of the restriction $\xi_L / \xi_L^{-1}(\mathrm{Cone}(L) - U) : \xi_L^{-1}(\mathrm{Cone}(L) - U) \to \mathrm{Cone}(L) - U$. Then the composition

$$\psi' = \psi\varphi / \bigl(\varphi^{-1}(\mathrm{Cone}(L) - U)\bigr) : \varphi^{-1}\bigl(\mathrm{Cone}(L) - U\bigr) \to \widetilde{K}$$

is a lifting of $\varphi / (\varphi^{-1}(\mathrm{Cone}(L) - U))$.

Next consider the set $\varphi^{-1}(\mathrm{cl}\, U) \subseteq B$ and $A = \varphi^{-1}(\mathrm{cl}\, U - U) \subseteq \varphi^{-1}(\mathrm{cl}\, U)$. Since $\mathrm{ed}(\varphi^{-1}(\mathrm{cl}\, U)) \leqslant [L]$, the composition

$$\varphi^{-1}(\mathrm{cl}\, U - U) \xrightarrow{\psi'} \mathrm{Cone}(L) \times L \xrightarrow{\mathrm{pr}_2} L$$

has an extension $\bar{\psi} : \mathrm{cl}\, U \to K$. Then the map $\bar{\varphi} : B \to \widetilde{L}$ defined by letting

$$\bar{\varphi}(b) = \begin{cases} \psi'(b), & \text{if } b \in \varphi^{-1}(\mathrm{Cone}(L) - U) \\ (\varphi(b), \bar{\psi}(b)), & \text{if } b \in \mathrm{cl}\, U, \end{cases}$$

is a lifting of φ. □

LEMMA 2.7. *Let L be a finite complex, X a (locally) finite polyhedron and $f : A \to L$ a simplicial map defined on a subpolyhedron A of X. Then there is an $[L]$-invertible approximately $[L]$-soft (proper) simplicial map $p = p_{(X,A,f)} : Y_{(X,A,f)} \to X$ satisfying the following conditions*:

(i) *$Y_{(X,A,f)}$ is a (locally) finite polyhedron.*

(ii) *The composition $fp/p^{-1}(A) : p^{-1}(A) \to L$ admits an extension over $Y_{(X,A,f)}$.*

PROOF. Extend f to a simplicial map $\bar{f} : X \to \mathrm{Cone}(L)$ and consider the diagram

$$\begin{array}{ccc} Y_{(X,A,f)} & \xrightarrow{\quad q \quad} & \widetilde{L} \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle \xi_L} \\ X & \xrightarrow{\quad \bar{f} \quad} & \mathrm{Cone}(L) \end{array}$$

where $Y_{(X,A,f)}$ is a pullback, i.e.,

$$Y_{(X,A,f)} = \{(x, z) \in X \times \widetilde{L}:\ \bar{f}(x) = \xi_L(z)\}$$

and the maps $p : Y_{(X,A,f)} \to X$ and $q : Y_{(X,A,f)} \to \widetilde{L}$ are restrictions onto $Y_{(X,A,f)}$ of the projections

$$X \times \widetilde{L} \to X \quad \text{and} \quad X \times \widetilde{L} \to \widetilde{L},$$

respectively. It is immediate that the map p is also $[L]$-invertible and approximately $[L]$-soft (it has same fibers as ξ_L). Next observe that $fp/p^{-1}(A) = \xi_L q/p^{-1}(A)$. Consequently the composition

$$Y_{(X,A,f)} \xrightarrow{q} \widetilde{L} \hookrightarrow \mathrm{Cone}(L) \times L \xrightarrow{\mathrm{pr}_2} L$$

is the desired extension of $fp/p^{-1}(A)$. □

Proposition 2.22, Lemma 2.7 and considerations of [49, 79] imply the following important result.

PROPOSITION 2.23. *Let L be a finite complex and X be a finite polyhedron or a compact Hilbert cube manifold. Then there exists an $[L]$-invertible and approximately $[L]$-soft map $f_X^{[L]} : \mu_X^{[L]} \to X$ satisfying the following conditions*:

(i) $\mu_X^{[L]} \in \mathrm{ANE}([L])$.

(ii) $\mathrm{ed}(\mu_X^{[L]}) = [L]$.

(iii) *For any map $f : B \to \mu_X^{[L]}$, where B is a compact space with $\mathrm{ed}(B) \leqslant [L]$, and for any open cover $\mathcal{U} \in \mathrm{cov}(\mu_X^{[L]})$ there is an embedding $g : B \to \nu_X^{[L]}$ which is $\mathcal{U}$-close to f and such that $gf_X^{[L]} = ff_X^{[L]}$.*

In particular, letting $X = \{pt\}$, we obtain the following statement.

THEOREM 2.5. *There exists a compact space $\mu^{[L]}$ satisfying the following conditions*:
(i) $\mu^{[L]}$ *is a compact* AE($[L]$)*-space.*
(ii) $\mathrm{ed}(\mu^{[L]}) = [L]$.
(iii) $\mu^{[L]}$ *is strongly* $[L]$*-universal for compact spaces.*

2.5.2. *Nonlocally compact case.* We start with the following two definitions.

DEFINITION 2.7. For a given complex L we say that a space X satisfies the *strong discrete approximation property modulo* L (notation: $[L]$-SDAP) if the set

$$\Bigl\{f \in C\Bigl(\bigoplus\{Y_k\colon k \in \omega\}, X\Bigr)\colon \bigl\{f(Y_k)\colon k \in \omega\bigr\} \text{ is discrete in } X\Bigr\},$$

where Y_k, $k \in \omega$, is a compactum such that $\mathrm{ed}(Y_k) \leqslant [L]$, is dense in the space $C(\bigoplus\{Y_k\colon k \in \omega\}, X)$.

DEFINITION 2.8. For a given complex L we say that a Polish space X is *strongly* $[L]$*-universal for Polish spaces* if for any Polish space B, its closed subspace A, any open cover $\mathcal{U} \in \mathrm{cov}(X)$ and any map $f : B \to X$ such that the restriction f/A is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.

COROLLARY 2.8. *Let L be a finitely dominated complex and X be a Polish* ANE($[L]$)*-space satisfying* $[L]$-SDAP. *Then the following conditions are equivalent for a closed subset A of X*:
(i) *A is a Z-set.*
(ii) *A is a strong Z-set.*

COROLLARY 2.9. *Let L be a finitely dominated complex. Every compact subset of a Polish* ANE($[L]$)*-space satisfying* $[L]$-SDAP *is a strong Z-set.*

PROPOSITION 2.24. *Let L be a finitely dominated complex and X be a Polish* ANE($[L]$)*-space. Then the following conditions are equivalent*:
(i) *X has the* $[L]$-SDAP.
(ii) *X is strongly* $[L]$*-universal for Polish spaces.*

PROOF. We prove only the absolute case. The relative version requires additional technical considerations.

Step 1. First we show that for each locally compact space Y with $\mathrm{ed}(Y) \leqslant [L]$ the set of closed embeddings of Y into X is a dense (and, according to Lemma 2.2, G_δ-) subset of the space $C(Y, X)$. Write $Y = Y_1 \cup Y_2$, where Y_1 and Y_2 are unions of discrete (in Y) collections of compacta. By (i) and Proposition 2.27, the set

$$C_i = \bigl\{f \in C(Y, X)\colon f(Y_i) \text{ is closed in } X\bigr\}$$

is dense and open in $C(Y, X)$, $i = 1, 2$. By Proposition 2.18, the set $D = C_1 \cap C_2$ is still open and dense and obviously consists of maps with closed in X images.

Next consider a countable open basis $\{U_k\colon k \in \omega\}$ of Y. Applying the above argument to each of the sets $Y - U_k$, we conclude that the set

$$D_k = \{f \in C(Y,X)\colon\ f(Y - U_k) \text{ is closed in } X\}$$

is open and dense in $C(Y,X)$, $k \in \omega$.

Since every open subset of Y is σ-compact, Corollary 2.9 and Proposition 2.18 imply that

$$T_{m,n} = \{f \in C(Y,X)\colon\ f(U_m) \cap f(U_n) = \emptyset\}$$

is a dense and G_δ-subset of $C(Y,X)$ for each $m,n \in \omega$ with $U_m \cap U_n = \emptyset$. Applying Proposition 2.18 once again we see that the intersection

$$D \cap \bigcap\{D_k\colon k \in \omega\} \cap \bigcap\{T_{m,n}\colon m,n \in \omega\}$$

is a dense and G_δ-subset of $C(Y,X)$. It only remains to note that this intersection consists of closed embeddings of Y into X.

Step 2. Now consider the general situation. Let $f\colon Y \to X$ be a map of a Polish space Y such that $\mathrm{ed}(Y) \leqslant [L]$ and let $\mathcal{U} \in \mathrm{cov}(Y)$. By Corollary 2.1, we may assume that Y is a closed subspace of a Polish AE($[L]$)-space $X^{[L]}$. Since X is an ANE($[L]$)-space, there is an extension $\tilde{f}\colon G \to X$ of the map f onto an open neighborhood G of Y in $X^{[L]}$. Let $\widetilde{\mathcal{U}}$ denote an open cover of G such that $St(\widetilde{\mathcal{U}})/X$ refines $\mathcal{U}$. By Proposition 2.11, G is a Polish ANE($[L]$)-space and, hence, by Lemma 2.5, there exist a locally compact space $K_{\widetilde{\mathcal{U}}}$ with $\mathrm{ed}(K_{\widetilde{\mathcal{U}}}) \leqslant [L]$ and two maps $\alpha\colon G \to K_{\widetilde{\mathcal{U}}}$ and $\beta\colon K_{\widetilde{\mathcal{U}}} \to G$ such that $\alpha\beta$ is $\widetilde{\mathcal{U}}$-close to id_G. Next consider a countable collection of compact subsets $\{A_k\colon k \in \omega\}$ of X with the following property: any closed subset A of X such that $A \subseteq X - \bigcup\{A_k\colon k \in \omega\}$ is a Z-set in X (the existence of such a collection follows from Proposition 2.19(i)). According to step 1 and Corollary 2.9, the map $\tilde{f}\beta\colon K_{\widetilde{\mathcal{U}}} \to X$ can be approximated by a closed embedding $\tilde{g}\colon K_{\widetilde{\mathcal{U}}} \to X$ such that $\tilde{g}(K_{\widetilde{\mathcal{U}}}) \cap \bigcup\{A_k\colon k \in \omega = \emptyset\}$. It is easy to see that the map $g = \tilde{g}\alpha\colon Y \to X$ approximates f, is a $\mathcal{U}$-map and $f(Y) \subseteq \bigcup\{A_k\colon k \in \omega\}$. This shows that the set of $\mathcal{U}$-maps of Y into X with images in $X - \bigcup\{A_k\colon k \in \omega\}$ is dense and (by Lemma 2.1) a G_δ-subset of $C(Y,X)$. Finally, Lemma 2.2 guarantees that the set of closed embeddings of Y into X with images in $X - \bigcup\{A_k\colon k \in \omega\}$ is dense and a G_δ-subset in $C(Y,X)$. By the choice of $\{A_k\colon k \in \omega\}$, every such embedding is a Z-embedding. □

THEOREM 2.6. *Let L be a finitely dominated complex and X be a Polish* A(N)E($[L]$)-*space. Then there exists an $[L]$-soft map $f_X^{[L]}\colon \nu_X^{[L]} \to X$ satisfying the following conditions*:

(i) $\nu_X^{[L]} \in$ A(N)E($[L]$).

(ii) $\mathrm{ed}(\nu_X^{[L]}) = [L]$.

(iii) *For any map $f\colon B \to \nu_X^{[L]}$, where B is a Polish space with* $\mathrm{ed}(B) \leqslant [L]$, *and for any open cover $\mathcal{U} \in \mathrm{cov}(\nu_X^{[L]})$, there is a* closed embedding $g\colon B \to \nu_X^{[L]}$ *which is $\mathcal{U}$-close to f and such that $gf_X^{[L]} = ff_X^{[L]}$.*

PROOF. Let $X_0 = X$. Let X_1 be a Polish space which admits an $[L]$-soft map $h_1 : X_1 \to X_0 \times R^\omega$ onto the product $X_0 \times R^\omega$ and $\mathrm{ed}(X_1) \leqslant [L]$ (the existence of X_1 and h_1 is guaranteed by Theorem 2.1). Let $p_0^1 = \pi_0 h_1$, where $\pi_0 : X_0 \times R^\omega \to X_0$ denotes the natural projection onto the first coordinate. In general, if a Polish space X_k with $\mathrm{ed}(X_k) \leqslant [L]$ has already been constructed, as the space X_{k+1} we take any Polish space which admits an $[L]$-soft map $h_{k+1} : X_{k+1} \to X_k \times R^\omega$ and such that $\mathrm{ed}(X_k) \leqslant [L]$. We also define the map p_k^{k+1} as the composition $p_k^{k+1} = \pi_k h_{k+1}$, where $\pi_k : X_k \times R^\omega \to X_k$ denotes the projection onto the first coordinate. Therefore we have an inverse sequence $\mathcal{S} = \{X_k, p_k^{k+1}, \omega\}$. Let $\nu_X^{[L]} = \lim \mathcal{S}$ and $f_X^{[L]} = p_0$. By Proposition 2.6, $\mathrm{ed}(\nu_X^{[L]}) \leqslant [L]$. Considerations similar to the proof of [51, Proposition 5.1.11] guarantee that the limit projection $f_X^{[L]} : \nu_X^{[L]} \to X$ has the desired properties. Observe also that $f_X^{[L]}$ (being the limit projection of the inverse sequence with $[L]$-soft short projections) is $[L]$-soft. In particular, by Proposition 2.13, $\nu_X^{[L]} \in \mathrm{A(N)E}([L])$ and $\mathrm{ed}(\nu_X^{[L]}) = [L]$. □

Letting $X = \{\mathrm{pt}\}$ in the above theorem we obtain the following important statement which significantly strengthens Corollary 2.1.

THEOREM 2.7. *Let L be a finitely dominated complex. Then there exists a space $\nu^{[L]}$ satisfying the following conditions*:
- (i) *$\nu^{[L]}$ is a Polish* $\mathrm{AE}([L])$*-space.*
- (ii) $\mathrm{ed}(\nu^{[L]}) = [L]$.
- (iii) *$\nu^{[L]}$ is strongly $[L]$-universal for Polish spaces.*

Different constructions (including direct geometrical ones) may produce at first glance different spaces satisfying the conditions of Theorem 2.7. For instance, it is not clear whether the spaces $\nu_{\mathrm{pt}}^{[L]}$ and $\nu_{[0,1]}^{[L]}$ are homeomorphic.

2.6. $[L]$-*homotopy*

We need to develop an adequate homotopy language for $\mathrm{ANE}([L])$-spaces. The key here is the following definition.

DEFINITION 2.9. Two (proper) maps $f_0, f_1 : X \to Y$ are said to be (*properly*) $[L]$-*homotopic* (notation: $f_0 \overset{[L]}{\simeq} f_1$, $f_0 \overset{[L]}{\simeq}_p f_1$) if for any (proper) map $h : Z \to X \times [0,1]$, where Z is a space with $\mathrm{ed}(Z) \leqslant [L]$, the composition $(f_0 \oplus f_1)h/h^{-1}(X \times \{0,1\}) : h^{-1}(X \times \{0,1\}) \to Y$ admits an (proper) extension $H : Z \to Y$.

EXAMPLE 2.7. Observe that:
- (i) $[L]$-homotopic maps are $[K]$-homotopic for each K with $[K] \leqslant [L]$.
- (ii) If X is an $\mathrm{AE}([L])$-space, then id_X is $[L]$-homotopic to a constant map.
- (iii) The concept of $[S^n]$-homotopic maps between complexes coincides with the notion of n-homotopic maps introduced by J.H.C. Whitehead in [151]. It should be emphasized that later, certain reasons within algebraic [15], [16, p. 124], [125]

and geometric [20, 43, 44, 46, 51] topology led to a shift of scale and Whitehead's $(n+1)$-types are now commonly referred as n-types. It seems now more appropriate to change the terminology back. Thus everywhere below, the n-homotopy denotes the S^n-homotopy, i.e., Whitehead's n-homotopy types.

PROPOSITION 2.25. *Let L and K be connected finite complexes and $[K] < [L]$. Then the identity map id_K is not $[L]$-homotopic to a constant map.*

PROOF. Consider an $[L]$-invertible and approximately $[L]$-soft map $f : X \to \mathrm{Cone}(K)$ where X is an AE($[L]$)-compactum with $\mathrm{ed}(X) = [L]$ (Proposition 2.23). Now, assuming that id_K is $[L]$-homotopic to a constant map, we see that there is a map $g : X \to K$ such that $g/f^{-1}(K) = f/f^{-1}(K)$ (here K is identified with the bottom, $K \times \{0\}$, of the cone $\mathrm{Cone}(K)$). Take a compactum Y such that $\mathrm{ed}(Y) = [L]$. Since $[K] < [L]$, there exists a map $\alpha : Y_0 \to K$, defined on a closed subspace Y_0 of Y, which does not admit an extension over Y. Since $\mathrm{ed}(Y_0) \leqslant [L]$ and f is $[L]$-invertible, there is a map $\beta : Y_0 \to X$ such that $\alpha = f\beta$. Since X is an AE($[L]$)-compactum and $\mathrm{ed}(Y) \leqslant [L]$, there is a map $\gamma : Y \to X$ such that $\beta = \gamma/Y_0$. Then the composition $g\gamma : Y \to K$ produces an extension of α. Contradiction. Thus id_K is not $[L]$-homotopic to a constant map. □

According to Proposition 2.7, the extension dimension of complexes coincides with the usual dimension. Nevertheless, as follows from the above discussion, the concept of $[L]$-homotopy differs from the classical notions of homotopy or n-homotopy, even for maps between complexes. Indeed, let $n \geqslant 1$ and L be a connected complex such that $[S^n] < [L] < [S^{n+1}]$ (Example 2.4(iii)). By Remark 2.1, Example 2.7(ii) and Proposition 2.25, we see that

$$\mathrm{id}_L \overset{[L]}{\simeq} \mathrm{const}, \text{ but } \mathrm{id}_L \overset{[S^{n+1}]}{\not\simeq} \mathrm{const}.$$

Similarly,

$$\mathrm{id}_{S^n} \overset{[S^n]}{\simeq} \mathrm{const}, \text{ but } \mathrm{id}_{S^n} \overset{[L]}{\not\simeq} \mathrm{const}.$$

Later we make use of some natural concepts associated with $[L]$-homotopy. Here are their definitions.

DEFINITION 2.10. A (proper) map $f : X \to Y$ is a *(proper) $[L]$-homotopy equivalence* if there is a (proper) map $g : Y \to X$ such that the compositions gf and fg are (properly) $[L]$-homotopic to id_X and id_Y, respectively.

DEFINITION 2.11. Let $\mathcal{U} \in \mathrm{cov}(Y)$. We say that maps $f, g : X \to Y$ are *$\mathcal{U}$-$[L]$-homotopic* if for any map $h : Z \to X \times [0,1]$, where Z is a space with $\mathrm{ed}(Z) \leqslant [L]$, the composition $(f_0 \oplus f_1)h/h^{-1}(X \times \{0,1\}) : h^{-1}(X \times \{0,1\}) \to Y$ admits an extension $H : Z \to Y$ such that the collection $\{H(h^{-1}(\{x\} \times [0,1])) : x \in X\}$ refines $\mathcal{U}$.

DEFINITION 2.12. A map $f : X \to Y$ is a *fine* $[L]$-*homotopy equivalence* if for any $\mathcal{U} \in \operatorname{cov}(Y)$ there is map $g_{\mathcal{U}} : Y \to X$ such that the composition $fg_{\mathcal{U}}$ is $\mathcal{U}$-$[L]$-homotopic to id_Y and the composition $g_{\mathcal{U}} f$ is $f^{-1}(\mathcal{U})$-$[L]$-homotopic to id_X.

2.6.1. *$[L]$-homotopy properties of* ANE($[L]$)*-spaces*

PROPOSITION 2.26. *Let L be a finitely dominated complex and $\mathcal{U} \in \operatorname{cov}(X)$ be an open cover of a Polish* ANE($[L]$)*-space X. Then there exists an open cover $\mathcal{V} \in \operatorname{cov}(X)$ such that any two $\mathcal{V}$-close maps of any space into X are $\mathcal{U}$-$[L]$-homotopic.*

PROOF. We may assume that X is a closed subspace of R^ω. Fix a proper $[L]$-invertible map $f : Y \to R^\omega$ such that $\operatorname{ed}(Y) \leqslant [L]$ (see Theorem 2.3). Since $X \in \mathrm{ANE}([L])$, the restriction $f/f^{-1}(X) : f^{-1}(X) \to X$ admits an extension $g : U \to X$ over a neighborhood U of $f^{-1}(X)$ in Y. The properness of f allows us to assume that $U = f^{-1}(V)$ for some neighborhood V of X in R^ω. Consider an open cover $\mathcal{U}' = \{V - f(g^{-1}(X - U))\colon U \in \mathcal{U}\}$ of V. Since V is a Polish ANE-space, there exists an open cover $\mathcal{V}' \in \operatorname{cov}(V)$ such that any two $\mathcal{V}'$-close maps into V are $\mathcal{U}'$-homotopic. Let[7] $\mathcal{V} = \mathcal{V}'/X$ and let us show that $\mathcal{V}$ is the desired open cover of X. Consider a space Z and any two $\mathcal{V}$-close maps $\alpha_0, \alpha_1 : Z \to X$. Then, by the construction, there is a homotopy $H : Z \times [0, 1] \to V$ such that $H(z, i) = \alpha_i(z)$ for each $(z, i) \in Z \times \{0, 1\}$ and the collection $\{H(\{z\} \times [0, 1])\colon z \in Z\}$ refines $\mathcal{U}'$. Now consider a map $\beta : \widetilde{Z} \to Z \times [0, 1]$ where $\widetilde{Z}$ is a space such that $\operatorname{ed}(\widetilde{Z}) \leqslant [L]$. Since $\operatorname{ed}(\widetilde{Z}) \leqslant [L]$, the $[L]$-invertibility of f guarantees the existence of a map $G : \widetilde{Z} \to U$ such that $fG = H\beta$. Finally, let $F = gG : \widetilde{Z} \to X$. One can easily verify that F is indeed an extension of $(\alpha_0 \oplus \alpha_1)\beta/\beta^{-1}(Z \times \{0, 1\})$ and that the collection $\{F(\beta^{-1}(\{z\} \times [0, 1]))\colon z \in Z\}$ refines $\mathcal{U}$. Thus α_0 and α_1 are $\mathcal{U}$-$[L]$-homotopic. □

The following two statements (their proofs follow the proof of Proposition 2.26) are versions of the "$[L]$-Homotopy Extension Theorem".

PROPOSITION 2.27. *Let L be a finitely dominated complex and X be a Polish* ANE($[L]$)*-space. Then for each $\mathcal{U} \in \operatorname{cov}(X)$, there exists $\mathcal{V} \in \operatorname{cov}(X)$ refining $\mathcal{U}$, such that the following condition holds*:

$(*)_L$ *For any space B with $\operatorname{ed}(B) \leqslant [L]$, any closed subspace A of B, and any two $\mathcal{V}$-close maps $f, g : A \to X$ such that f has an extension $F : B \to X$, it follows that g also has an extension $G : B \to X$ which is $\mathcal{U}$-$[L]$-homotopic to F.*

PROPOSITION 2.28. *Let L be a finitely dominated complex and X be a Polish* ANE($[L]$)*-space. Suppose that A is closed in a space B with $\operatorname{ed}(B) \leqslant [L]$. If maps $f, g : A \to X$ are $[L]$-homotopic and f admits an extension $F : B \to X$, then g also admits an extension $G : B \to X$, and it may be assumed that $F \overset{[L]}{\simeq} G$.*

DEFINITION 2.13. We say that a compactum X is a $UV^{[L]}$-*compactum* if the following condition is satisfied:

[7] $\mathcal{V}'/X = \{V' \cap X\colon V' \in \mathcal{V}'\}$.

– For any embedding of X into an ANE($[L]$)-space Y with $\text{ed}(Y - X) \leqslant [L]$, for any neighborhood U of X in Y there is a smaller neighborhood V such that the inclusion $V \hookrightarrow U$ is $[L]$-homotopic (in U) to a constant map.

Not surprisingly, we implicitly have already met $UV^{[L]}$-compacta. Indeed, the fibers of proper approximately $[L]$-soft maps (between ANE($[L]$)-spaces) are $UV^{[L]}$-compacta. This suggests another, perhaps more familiar, name for approximately $[L]$-soft maps – $UV^{[L]}$-maps. The standard argument [102] proves the following statement.

PROPOSITION 2.29. *Let $X \to Y$ be a map between* ANE($[L]$)*-compacta and* $\text{ed}(Y) \leqslant [L]$. *Then the following conditions are equivalent*:

(i) *f is a $UV^{[L]}$-map.*
(ii) *f is approximately $[L]$-soft.*
(iii) *f is a fine $[L]$-homotopy equivalence.*

The following result provides an internal characterization of ANE($[L]$)-spaces. Its proof is standard (see [24] for details).

THEOREM 2.8. *Let $[L] \leqslant [S^n]$. Then the following conditions are equivalent for any space X*:

(i) *X is an* ANE($[L]$)*-space.*
(ii) *X is* locally $[L]$-contractible, *i.e., for each neighborhood U_x of any point $x \in X$, there exists a smaller neighborhood V_x such that the inclusion $V_x \hookrightarrow U_x$ is $[L]$-homotopic (in U_x) to a constant map.*

How essential is the requirement $[L] \leqslant [S^n]$? What happens if the extension type $[L]$ is larger than any $[S^n]$ (which, according to Example 2.1(iv), means that $[L] = [\{\text{pt}\}]$) or is not comparable with any $[S^n]$ (this possibility, according to Example 2.5, also can be realized)? In both cases the answer is the same – the statement in these cases is not true. Borsuk's example [24] of a locally contractible space which is not an ANE-space serves as a counterexample in both cases.

2.6.2. *$[L]$-homotopy groups.* The next step in our program is to define algebraic $[L]$-homotopy invariants. Here we discuss only $[L]$-homotopy groups. Everywhere below L stands for a connected finite complex.

For each $n \geqslant 0$ we consider an *"nth $[L]$-sphere"* $S^n_{[L]}$, which is nothing else but an $[L]$-dimensional ANE($[L]$)-compactum ($\mu^{[L]}_{S^n}$ in the notations of Proposition 2.23) admitting an approximately $[L]$-soft map onto S^n. It can be shown that all possible choices of an $[L]$-sphere $S^n_{[L]}$ are $[L]$-homotopy equivalent. This allows us to consider, for each $n \geqslant 1$ and for any space X, the set $\pi^{[L]}_n(X) = [S^n_{[L]}, X]_{[L]}$ of $[L]$-homotopy classes of maps of the $[L]$-sphere $S^n_{[L]}$ into X. It turns out that this set carries a natural group structure, i.e., for an $[L]$-homotopy class $[f]_{[L]} \in [X, Y]_{[L]}$, the natural map

$$\pi^{[L]}_n([f]_{[L]}) : \pi^{[L]}_n(X) \to \pi^{[L]}_n(Y)$$

is a well defined homomorphism. We call the group $\pi_n^{[L]}(X)$ the *nth* $[L]$*-homotopy group* of the space X. Even for a complex K, the system $\{\pi_n^{[L]}(K)\colon n = 1, 2, \ldots\}$ of $[L]$-homotopy groups may differ from the system $\{\pi_n(K)\colon n = 1, 2, \ldots\}$ of usual homotopy groups.

EXAMPLE 2.8. (i) If $[L] = [\{\mathrm{pt}\}]$, then the kth $[\{\mathrm{pt}\}]$-sphere is (homotopically) the same as the usual sphere S^k, and consequently the system of $[\{\mathrm{pt}\}]$-homotopy groups of any space X coincides with the system of usual homotopy groups

$$\pi_1(X), \pi_2(X), \ldots.$$

(ii) If $[L] = [S^n]$, then kth $[S^n]$-sphere $S^k_{[S^n]}$ is the same (homotopically) as the usual sphere S^k precisely for $k \leqslant n-1$. For $k \geqslant n$, the kth $[S^n]$-sphere $S^k_{[S^n]}$ is $[S^n]$-homotopy equivalent to S^k, which in turn (since $k \geqslant n$) is $[S^n]$-homotopy equivalent to a point. Therefore the system of $[S^n]$-homotopy groups of a space X looks as follows

$$\pi_1(X), \pi_2(X), \ldots, \pi_{n-1}(X), 0, 0, \ldots, 0, \ldots.$$

What can we say about maps $f\colon X \to Y$ which induce homomorphisms of *all* $[L]$-homotopy groups? Let us consider only spaces with nice local $[L]$-homotopy structure. For instance, consider ANE($[L]$)-spaces (see Theorem 2.8) of extension dimension not exceeding $[L]$. In addition, let us assume (in order to avoid technical difficulties which are not essential for the present discussion ... see Problem 2.14) that both X and Y are $[L]$-homotopy equivalent to spaces which admit (proper) approximately $[L]$-soft and $[L]$-invertible maps onto locally finite polyhedra. Let us temporarily call such spaces triangulable ANE($[L]$)-spaces or simply $[L]$-complexes.

THEOREM 2.9. *A map between $[L]$-complexes is an $[L]$-homotopy equivalence if and only if it induces isomorphisms of* all *$[L]$-homotopy groups.*

What does this theorem tell us in the cases $[L] = [\{\mathrm{pt}\}]$ and $[L] = [S^n]$?

If $[L] = [\{\mathrm{pt}\}]$, then $[L]$-complexes are precisely (locally compact) ANE-spaces (this follows from Theorems 3.34 and 3.39 or 3.22 and 3.27). Therefore, according to Example 2.8(i) we obtain Whitehead's well known result.

THEOREM 2.10 ([150, 151]). *A map between* ANE*-spaces is a homotopy equivalence if and only if it induces isomorphisms of all homotopy groups.*

There is another important result of Whitehead which states the following.

THEOREM 2.11 ([150, 151]). *A map between at most n-dimensional* ANE(n)*-spaces is an n-homotopy equivalence if and only if it induces isomorphisms of the kth homotopy groups for each $k \leqslant n-1$.*

Why are only the first $n-1$ homotopy groups responsible for making a map an n-homotopy equivalence? What is common between Theorems 2.10 and 2.11? How is the last result related to Theorem 2.9? A crucial observation here is that the system of the first $n-1$ homotopy groups is the initial segment of the system of *all* $[S^n]$-homotopy groups, the remaining part of which *always* consists of trivial groups (Example 2.8(ii)) and therefore cannot be seen explicitly. Remarking now that $[S^n]$-complexes are precisely at most n-dimensional LC^{n-1}-spaces (this follows from the results of Section 4), we arrive at the following corollary of Theorem 2.9, which is just a reformulation of Theorem 2.11 but allows us to understand the whole picture.

THEOREM 2.12. *A map between at most n-dimensional LC^{n-1}-spaces is an n-homotopy equivalence if and only if it induces isomorphisms of* all *$[S^n]$-homotopy groups.*

It is easy to see that the category of complexes and $[S^1]$-types is isomorphic to the category of sets and their maps. It is known [114, 151] that:

(a) The category of connected complexes and their $[S^2]$-types can be identified with the category of groups and their homomorphisms.
(b) The category of connected complexes and their $[S^3]$-types is isomorphic to the category of *crossed modules*.

For $n \geqslant 4$ there are partial results characterizing the category of complexes and $[S^n]$-types. See [15] for further information.

In the light of these results, the problem of the algebraic characterization of the *$[L]$-homotopy category* $\mathrm{HOMOT}_{[L]}$ becomes very intriguing.

PROBLEM 2.3. Characterize algebraically the category of $[L]$-complexes and their $[L]$-homotopy types.

In the case $L = \{\mathrm{pt}\}$ there is an important notion of *simple homotopy* (see [62] for details). It was shown in [37] that a map $f : X \to Y$ between compact ANE-spaces is a simple homotopy equivalence if and only if there exist a compact ANE-space Z and two cell-like maps $\alpha : Z \to X$ and $\beta : Z \to Y$ such that $f\alpha \simeq \beta$. In the case $L = S^n$ there are no simple homotopy obstructions. We conjecture that the same applies to complexes L with $[L] \leqslant [S^n]$. In the remaining cases, i.e., for complexes the extension types of which are not comparable with any $[S^n]$, the situation is not clear. Of course, guided by the above cited result of Chapman, we can formally introduce the concept of simple $[L]$-homotopy equivalence.

DEFINITION 2.14. A (proper) map $f : X \to Y$ between ANE($[L]$)-spaces is called a (proper) *simple $[L]$-homotopy equivalence* if there exist an ANE($[L]$)-space Z and (proper) $UV^{[L]}$-maps $\alpha : Z \to X$ and $\beta : Z \to Y$ such that $f\alpha$ is (properly) $[L]$-homotopic to β.

We specifically ask:

PROBLEM 2.4. Let L be a finite complex not comparable with any $[S^n]$. Do the classes of $[L]$-homotopy equivalences and simple $[L]$-homotopy equivalences differ? More specifically, is it true that there exists an $[L]$-homotopy equivalence between ANE($[L]$)-compacta of extension dimension not exceeding $[L]$ which is not a simple $[L]$-homotopy equivalence?

2.7. $[L]$-*shape*

In this section we define the $[L]$-*shape category* (notation: $\mathrm{SHAPE}_{[L]}$). In the classical case (i.e., $L = \{\mathrm{pt}\}$) there are at least three alternative approaches to the concept of shape: geometrical, spectral and categorical (see [25, 84, 115] for complete details). There is also the fourth approach based on Chapman's complement theorem [33]. We will not discuss algebraic aspects of the $[L]$-shape theory, and instead will concentrate on its geometrical side. Although it is better to have the $[L]$-shape category defined over the class of all compacta, for technical reasons we prefer to discuss here the category $\mathrm{SHAPE}_{[L]}$ whose objects are compacta of extension dimension not exceeding $[L]$. Morphisms $\alpha \in \mathrm{Mor}(X, Y)$ of this category are defined as maps[8]

$$\alpha : \bigcup \{[Y, P]_{[L]} \colon\ P \in \mathrm{ANE}([L]) \text{ and } \mathrm{ed}(P) \leqslant [L]\} \to \bigcup \{[X, P]_{[L]} \colon\ P \in \mathrm{ANE}([L]) \text{ and } \mathrm{ed}(P) \leqslant [L]\}$$

satisfying the following two conditions:

(i) If $P \in \mathrm{ANE}([L])$ and $\mathrm{ed}(P) \leqslant [L]$, then $\alpha([Y, P]_{[L]}) \subseteq [X, P]_{[L]}$.

(ii) If $P, P' \in \mathrm{ANE}([L])$, $\mathrm{ed}(P) \leqslant [L]$, $\mathrm{ed}(P') \leqslant [L]$, $\psi \in [Y, P]_{[L]}$, $\psi' \in [Y, P']_{[L]}$ and $\varphi \in [P, P']_{[L]}$ are such that $\varphi\psi = \psi'$, then $\varphi\alpha(\psi) = \alpha(\psi')$.

Next we define the $[L]$-*fundamental functor* $\mathcal{F}_{[L]} : \mathrm{HOMOT}_{[L]} \to \mathrm{SHAPE}_{[L]}$. It acts as follows: $\mathcal{F}_{[L]}(X) = X$ for any compactum X with $\mathrm{ed}(X) \leqslant [L]$; if $\varphi \in [X, Y]_{[L]}$, then the $[L]$-shape morphism $\mathcal{F}_{[L]}(\varphi) : X \to Y$ is given by the rule: $\mathcal{F}_{[L]}(\psi) = \psi\varphi$ for each $\psi \in [Y, P]_{[L]}$, where P is an arbitrary $\mathrm{ANE}([L])$-compactum with $\mathrm{ed}(P) \leqslant [L]$. The composition of the functors $\mathcal{H}_{[L]}$ (the natural $[L]$-*homotopy functor*) and $\mathcal{F}_{[L]}$ is denoted by $\mathcal{S}_{[L]} : \mathrm{COMP}([L]) \to \mathrm{SHAPE}_{[L]}$ and called the $[L]$-*shape functor* (here $\mathrm{COMP}([L])$ denotes the class of compacta whose extension dimension does not exceed $[L]$). We now arrive at the natural definition.

DEFINITION 2.15. Let X and Y be compacta with $\mathrm{ed}(X), \mathrm{ed}(Y) \leqslant [L]$. We say that X and Y have the same $[L]$-shape (and write $\mathrm{Sh}_{[L]}(X) = \mathrm{Sh}_{[L]}(Y)$) if X and Y are isomorphic objects of the category $\mathrm{SHAPE}_{[L]}$.

Obviously $[L]$-homotopy equivalent compacta have the same $[L]$-shape but not vise versa. It is also an easy exercise to show that $\mathrm{Sh}_{[L]}(X) = \mathrm{Sh}_{[L]}(\{\mathrm{pt}\})$ if and only if X is an $UV^{[L]}$-compactum.

PROBLEM 2.5. A map $f : X \to Y$ is a *hereditary* $[L]$-*shape equivalence* if for each closed subspace A of Y, the $[L]$-shape morphism generated by the map $f/f^{-1}(A) : f^{-1}(A) \to A$ is an $[L]$-shape equivalence. Do the classes of $UV^{[L]}$-maps and hereditary $[L]$-shape equivalences between compacta of extension dimension $\leqslant [L]$ coincide?

Is it possible to define a concept of $[L]$-shape dimension? How about algebraic $[L]$-shape invariants? Is there a satisfactory approach to the notion of strong $[L]$-shape? Another interesting question:

[8] $[Y, P]_{[L]}$ stands for the set of $[L]$-homotopy classes of maps of Y into P.

PROBLEM 2.6. Characterize $[L]$-dimensional compacta with a polyhedral $[L]$-shape.

Perhaps the most fundamental geometrical property of $[L]$-shapes is expressed in the following result (the cases $L = \{\mathrm{pt}\}$ and $L = S^n$ were considered in [33] and in [43, 44], respectively).

THEOREM 2.13 (Complement Theorem for $\mu^{[L]}$). *Let X and Y be Z-sets in the Menger compactum $\mu^{[L]}$ modulo L. Then the following conditions are equivalent:*

(i) *The complements $\mu^{[L]} - X$ and $\mu^{[L]} - Y$ are homeomorphic.*

(ii) $\mathrm{Sh}_{[L]}(X) = \mathrm{Sh}_{[L]}(Y)$, *i.e., X and Y have the same $[L]$-shape.*

The proof of this theorem can be carried out under the assumption of the validity of Conjecture 2.8.

In fact, the Complement Theorem has a categorical nature. In order to formulate the precise result we first need to introduce the corresponding definitions. Loosely speaking, two proper maps $f, g : X \to Y$ between ANE($[L]$)-spaces of extension dimension $\leqslant [L]$ are said to be *weakly properly $[L]$-homotopic* if for each compactum $B \subseteq Y$ there is a compactum $A \subseteq X$ such that the restrictions $f/(X - A)$ and $g/(X - A)$ are properly $[L]$-homotopic (in the sense of Definition 2.9) in $Y - B$. Now consider the full subcategory $\mathcal{A}$ of the category $\mathrm{SHAPE}_{[L]}$ whose objects are Z-sets of the compactum $\mu^{[L]}$. Also consider the category $\mathcal{B}$ whose objects are complements of Z-sets in $\mu^{[L]}$ and whose morphisms are weak proper $[L]$-homotopy classes of proper maps.

PROPOSITION 2.30. *There is a categorical isomorphism $\mathcal{T} : \mathcal{A} \to \mathcal{B}$ such that $\mathcal{T}(X) = \mu^{[L]} - X$ for each $X \in \mathcal{A}$.*

Complements of Z-sets in $\mu^{[L]}$ are $\mu^{[L]}$-manifolds (under the assumption of Conjecture 2.7) which have $\mu^{[L]}$-manifold compactifications with "boundaries" ($\mu^{[L]}$ being the compactification and Z-sets being boundaries). The following question has its roots in [28, 38, 39, 50, 57, 131].

PROBLEM 2.7. When is it possible, for a given $\mu^{[L]}$-manifold X, to find a $\mu^{[L]}$-manifold compactification $\widetilde{X}$ such the complement (i.e., "boundary") $\widetilde{X} - X$ is a Z-set in $\widetilde{X}$? When can the boundary $\widetilde{X} - X$ be assumed to be a $\mu^{[L]}$-manifold?

The last part of the above question might be closely related to the problem of the recognition of spaces of finite $[L]$-homotopy type. In order to formulate the question, we need to introduce the class of *finitely $[L]$-homotopy dominated spaces*. A space X is said to be $[L]$-homotopy dominated by a space Y if there are two maps $f : X \to Y$ and $g : Y \to X$ such that

$$gf \overset{[L]}{\simeq} \mathrm{id}_X .$$

If there exists an ANE($[L]$)-compactum Y such that X is $[L]$-homotopy dominated by Y and $\mathrm{ed}(Y) \leqslant [L]$, then we say that X is finitely $[L]$-homotopy dominated. If X is

$[L]$-homotopy equivalent to an ANE($[L]$)-compactum Y with $\text{ed}(Y) \leqslant [L]$ then X has a *finite* $[L]$*-homotopy type*. If $[L] = [\{\text{pt}\}]$, then the well-known result of Wall [144] states that there are algebraic obstructions for a finitely homotopy dominated ANE-space to have a finite homotopy type. In the case $[L] = [S^n]$, the analog of Wall's obstruction vanishes, and consequently every finitely $[S^n]$-homotopy dominated ANE($[S^n]$)-spaces has a finite $[S^n]$-homotopy type [50]. In light of these results, we conjecture the following.

CONJECTURE 2.1. Let L be a finite complex. If $[L] \leqslant [S^n]$, then every finitely $[L]$-homotopy dominated locally compact ANE($[L]$)-space X with $\text{ed}(X) \leqslant [L]$ has a finite $[L]$-homotopy type. If $[L]$ is not comparable with any $[S^n]$, then there are finitely $[L]$-homotopy dominated locally compact ANE($[L]$)-spaces X with $\text{ed}(X) \leqslant [L]$ which do not have a finite $[L]$-homotopy type.

In the last case it would be very interesting to find explicit constructions of the corresponding algebraic obstructions.

2.8. *Open problems*

In this subsection we discuss possible avenues of development of infinite dimensional topology modulo a *finitely complex*. As in the classical cases $L = \{\text{pt}\}$, $L = S^n$ the main objects of investigation are the above introduced spaces $\mu^{[L]}$, $\nu^{[L]}$ and manifolds (shortly $\mu^{[L]}$- and $\nu^{[L]}$-manifolds) modeled on these spaces.

PROBLEM 2.8. Find direct geometric constructions of the spaces $\mu^{[L]}$ and $\nu^{[L]}$.

The questions below implicitly follow the so-called Edwards' strategy. This is the sequence of steps which proved to be successful in the cases of I^{ω}-, R^{ω}- and μ^n-manifolds and there is no reason to believe that it will not work in the general situation. Here are its main ingredients. Suppose we have a model space X (in our case the spaces $\nu^{[L]}$ and $\mu^{[L]}$) and we would like to develop a theory of X-manifolds. Suppose also that we already have a plausible conjecture about the topological characterization of X-manifolds. The first step of the strategy is to "resolve" a space Y, satisfying the conditions of the conjecture, via a "nice" resolution map $f_Y : X_Y \to Y$ defined on an X-manifold. The second step uses Bing's shrinking criterion (Theorem 2.4) and establishes that the map f_Y is actually a near-homeomorphism. Of course, there are several other essential (and, as a rule, hard to prove) components of this program hidden in each of the indicated steps. Almost all of them are discussed in the subsequent questions. But going back to the first step, a natural question arises: what are those nice maps? In the case of $\nu^{[L]}$- and $\mu^{[L]}$-manifolds such maps have already been introduced – these are approximately $[L]$-soft maps. Therefore our first questions are not unexpected.

CONJECTURE 2.2 ($\nu^{[L]}$*-manifold resolution*). Polish A(N)E($[L]$)-spaces are approximately $[L]$-soft images of $\nu^{[L]}$-manifolds.

CONJECTURE 2.3 (*$\mu^{[L]}$-manifold resolution*). Locally compact A(N)E($[L]$)-spaces are proper approximately $[L]$-soft images of $\mu^{[L]}$-manifolds.

The following two questions correspond to the second step of the Edwards' program.

CONJECTURE 2.4 (*Near-homeomorphisms of $\nu^{[L]}$-manifolds*). The class of approximately $[L]$-soft maps $f:\nu_X^{[L]} \to X$ of a $\nu^{[L]}$-manifold $\nu_X^{[L]}$ into a Polish A(N)E($[L]$)-space with the [L]-SDAP and of extension dimension $\mathrm{ed}(X) = [L]$ coincides with the class of near-homeomorphisms.

CONJECTURE 2.5 (*Near-homeomorphisms of $\mu^{[L]}$-manifolds*). A proper approximately $[L]$-soft map $f:\mu_X^{[L]} \to X$ of a $\mu^{[L]}$-manifold $\nu_X^{[L]}$ onto a locally compact A(N)E($[L]$)-space with the [L]-DDP and of extension dimension $\mathrm{ed}(X) = [L]$ is a near-homeomorphism.

The above conjectures immediately imply topological characterizations of the model spaces.

CONJECTURE 2.6 (*Topological characterization of $\nu^{[L]}$-manifolds*). A Polish space X is a $\nu^{[L]}$-manifold if and only if it has the following properties:

(i) $X \in \mathrm{ANE}([L])$.
(ii) $\mathrm{ed}(X) = [L]$.
(iii) X is strongly $[L]$-universal for Polish spaces or, equivalently, X has the [L]-SDAP.

CONJECTURE 2.7 (*Topological characterization of $\mu^{[L]}$-manifolds*). A locally compact space X is a $\mu^{[L]}$-manifolds if and only if it has the following properties:

(i) $X \in \mathrm{ANE}([L])$.
(ii) $\mathrm{ed}(X) = [L]$.
(iii) X is strongly $[L]$-universal for (locally) compact spaces or, equivalently, X has the [L]-DDP.

Z-set unknotting theorems would apparently play an essential role in attempts to prove the above conjectures.

CONJECTURE 2.8 (*Global Z-set unknotting*). Let $Y = \{\nu^{[L]}, \mu^{[L]}\}$ and let $h: A \to B$ be a homeomorphism between Z-sets of a Y-manifold X such that $h \overset{[L]}{\simeq} i_A$, where $i_A : A \hookrightarrow X$ denotes the inclusion map. Then there exists a homeomorphism $H: X \to X$ such that $h = H/A$ and $H \overset{[L]}{\simeq} \mathrm{id}_X$.

CONJECTURE 2.9 (*Local Z-set unknotting*). Let $Y = \{\nu^{[L]}, \mu^{[L]}\}$ and let $\mathcal{U} \in \mathrm{cov}(X)$ be an open cover of a Y-manifold X. Then there exists an open cover $\mathcal{V} \in \mathrm{cov}(X)$ such that the following condition is satisfied:

($*$) if $h: A \to B$ is a homeomorphism between Z-sets of X which is $\mathcal{V}$-close to i_A, then there exists a homeomorphism $H: X \to X$ which is $\mathcal{U}$-close to id_X and which extends h.

The possibility of classification of $\mu^{[L]}$- and $\nu^{[L]}$-manifolds by their $[L]$-homotopy types seems natural. Standard examples of $[L]$-homotopy equivalences are approximately $[L]$-soft maps between ANE($[L]$)-spaces of extension dimension not exceeding $[L]$.

CONJECTURE 2.10 (*Homotopy classification of $\nu^{[L]}$-manifolds*). $[L]$-homotopy equivalent $\nu^{[L]}$-manifolds are homeomorphic. More precisely, an $[L]$-homotopy equivalence of $\nu^{[L]}$-manifolds is $[L]$-homotopic to a homeomorphism.

The situation in the locally compact case is not expected to be so simple. First of all, we have to consider proper $[L]$-homotopy types. Secondly, even in the compact case we might face obstructions of a different nature (for instance, in the case $L = \{\mathrm{pt}\}$, simple homotopies have to be considered; see Section 3.2).

CONJECTURE 2.11 (*Simple homotopy classification of $\mu^{[L]}$-manifolds*). Let L be a finite complex such that $[L]$ is not comparable with any $[S^n]$. Then simply $[L]$-homotopy equivalent $\mu^{[L]}$-manifolds are homeomorphic. Moreover, every simple $[L]$-homotopy equivalence between $\mu^{[L]}$-manifolds is $[L]$-homotopic to a homeomorphism.

For a given $\mu^{[L]}$-manifold X, consider a Z-embedding $i : X \to X$ which is properly $[L]$-homotopic to id_X. The complement $X - i(X)$ will be called the *$[L]$-homotopy kernel* of X and will be denoted by $\mathrm{Ker}_{[L]}(X)$. This space is well defined if global Z-set unknotting holds in X. One would think of $\mathrm{Ker}_{[L]}(X)$ as an analog of the product $X \times [0, 1)$ in the category of $\mu^{[L]}$- manifolds. It is a simple observation that $\mu^{[L]}$-manifolds with homeomorphic (even $[L]$-homotopy equivalent) $[L]$-homotopy kernels have the same $[L]$-homotopy type. We conjecture the converse.

CONJECTURE 2.12 (*Homotopy classification of $\mu^{[L]}$-manifolds*). $[L]$-homotopy equivalent $\mu^{[L]}$-manifolds have homeomorphic $[L]$-homotopy kernels.

CONJECTURE 2.13 (*Triangulation of $\nu^{[L]}$-manifolds*). Every $\nu^{[L]}$-manifold is triangulable, i.e., it admits an $[L]$-soft map onto a locally finite polyhedron.

CONJECTURE 2.14 (*Triangulation of $\mu^{[L]}$-manifolds*). Every $\mu^{[L]}$-manifold is triangulable, i.e., it admits a proper approximately $[L]$-soft (and $[L]$-invertible) map onto a locally finite polyhedron.

The property of being a $\nu^{[L]}$- or $\mu^{[L]}$-manifold is hereditary with respect to open subspaces (assuming that Conjectures 2.6 and 2.7 are true). We conjecture that there are no other $\nu^{[L]}$- or $\mu^{[L]}$-manifolds.

CONJECTURE 2.15 (*Open embeddings of $\nu^{[L]}$-manifolds*). Every $\nu^{[L]}$-manifold admits an open embedding into $\nu^{[L]}$.

CONJECTURE 2.16 (*Open embeddings of $\mu^{[L]}$-manifolds*). The $[L]$-homotopy kernel of a $\mu^{[L]}$-manifold admits an open embedding into $\mu^{[L]}$.

Perhaps the following question should be asked first.

PROBLEM 2.9 (*Topological homogeneity*). Are the spaces $\nu^{[L]}$ and $\mu^{[L]}$ topologically homogeneous? For what L are these spaces topological groups?

Note that if L is not connected, then $\mu^{[L]}$ and $\nu^{[L]}$, being the Cantor set and the space of irrationals, are both topological groups.

Ideally one would expect that the spaces $\nu^{[L]}$ and $\mu^{[L]}$ are related to each other. In the known particular examples $\nu^{[L]}$ can be identified with the subspace of the compactum $\mu^{[L]}$, the complement $\mu^{[L]} - \nu^{[L]}$ of which is *locally* $[L]$-*homotopy negligible* in $\mu^{[L]}$. The last concept is defined as follows.

DEFINITION 2.16 ([139]). A subspace Y of a space X is locally $[L]$-homotopy negligible if for each open subset U of X the inclusion $U - Y \hookrightarrow U$ is an $[L]$-homotopy equivalence.

Now consider the compactum $\mu^{[L]}$ and fix a metric ρ on it. Theorem 2.5 allows us to construct a sequence $\{g_n : \mu^{[L]} \to \mu^{[L]}\}$ of Z-embeddings satisfying the following conditions:

(i) $\rho(\mathrm{id}_{\mu^{[L]}}, g_n) < 2^{-n}$.

(ii) $g_{n+1}/g_n(\mu^{[L]}) = \mathrm{id}_{g_n(\mu^{[L]})}$.

It turns out that the union

$$\Sigma^{[L]} = \bigcup\{g_n(\mu^{[L]}) \colon n \in \omega\}$$

is locally $[L]$-homotopy negligible in $\mu^{[L]}$ and its complement $\mu^{[L]} - \Sigma^{[L]}$ is a Polish AE($[L]$)-space of extension dimension $[L]$ with the [L]-SDAP. In other words, if Conjecture 2.6 is true, then $\nu^{[L]} \approx \mu^{[L]} - \Sigma^{[L]}$.

This observation suggests an alternative approach (through compactifications) to the solution of Conjecture 2.6 under the assumption that Conjecture 2.7 is true. Namely if a space X, satisfying the conditions of Conjecture 2.6, has, in addition, a compactification $\widetilde{X}$ which itself is an ANE($[L]$)-space, $\mathrm{ed}(\widetilde{X}) = [L]$ and the complement $\widetilde{X} - X$ is locally $[L]$-homotopy negligible, then it can easily be seen that $\widetilde{X}$ is strongly $[L]$-universal for compact spaces. Therefore, under our assumption, $\widetilde{X}$ is a $\mu^{[L]}$-manifold. This explains why we ask the following question.

PROBLEM 2.10. Let X be a space satisfying the conditions of Conjecture 2.6. Does there exist a $\mu^{[L]}$-manifold compactification $\widetilde{X}$ of X such that $\widetilde{X} - X$ is locally $[L]$-homotopy negligible in $\widetilde{X}$?

It should be noted here that this question remains open even in the case $L = S^n$. We remark also that in the general theory of ANE($[L]$)-spaces, the class of Polish spaces plays the central role. In part this is because every ANE($[L]$)-space (of extension dimension $[L]$) has an ANE($[L]$)-completion (of the same extension dimension) with locally $[L]$-homotopy negligible complement (see [53, Proposition 5.9, Case $X = \{\mathrm{pt}\}$]). The further reduction from Polish spaces to compact spaces is not always possible to make.

The subset $\Sigma^{[L]}$ of $\mu^{[L]}$ is actually a Z-skeletoid (or, alternatively, an absorber) in $\mu^{[L]}$. There are several possible approaches (such as Anderson's cap sets [32], Bessaga–Pełczyński's skeletoids [18, 19], West's absorbers [147], Geoghegan–Summerhill's pseudo-boundaries [98], Bestvina–Mogilski's absorbing sets [22], Baars–Gladdines–van Mill's absorbing systems [11]) to the concept of an absorber (see [14] for a complete discussion of the theory of absorbing sets). Further investigation of properties of the space $\Sigma^{[L]}$ gives rise the following conjecture.

CONJECTURE 2.17 (*Topological characterization of* $\Sigma^{[L]}$). The space $\Sigma^{[L]}$ is topologically the only space satisfying the following conditions:

(i) $\Sigma^{[L]}$ is a σ-compact AE($[L]$)-space.

(ii) $\mathrm{ed}(\Sigma^{[L]}) = [L]$.

(iii) $\Sigma^{[L]}$ is strongly $[L]$-universal with respect to σ-compact spaces.

(iv) $\Sigma^{[L]}$ has the [L]-SDAP or, equivalently, $\Sigma^{[L]}$ is the countable union of its strong Z-sets.

The spaces $\mu^{[L]}$, $\nu^{[L]}$ and $\Sigma^{[L]}$ are the representatives of the first three classes of the Borel hierarchy. A natural question arises of how to define and investigate strongly $[L]$-universal (in the corresponding sense) AE($[L]$)-spaces of extension dimension $[L]$ in higher Borel classes (see [14, 22] for further details in the case $[L] = \{\mathrm{pt}\}$).

3. The case $L = \{\mathrm{pt}\}$: "genuine" infinite dimensional topology

In this section we consider the case $L = \{\mathrm{pt}\}$. We will not discuss any statements presented in Section 2 for the general situation unless there are important additions and comments to be made. We use the terminology adapted earlier, but omit extra notational details. For instance in this section ANE-spaces mean ANE($[\{\mathrm{pt}\}]$)-spaces, soft maps are same as $[\{\mathrm{pt}\}]$-soft maps etc.

The following statement [23, 81] provides a basic example of absolute extensors.

THEOREM 3.1. *If C is a convex subspace of a locally convex linear topological space, then $C \in$ AE(X) for each metrizable space X.*

The local convexity in the above statement is essential. The corresponding example has been constructed in [30].

THEOREM 3.2 ([109]). *Every infinite dimensional compact convex subspace of an arbitrary Fréchet space is homeomorphic to the Hilbert cube.*

THEOREM 3.3 ([71]). *The following conditions are equivalent for a nonlocally compact Polish group X:*

(i) *X is an* ANE-*space* (*respectively,* AE-*space*).

(ii) *X is an R^{ω}-manifold* (*respectively, is homeomorphic to R^{ω}*).

Theorems 3.1 and 3.3 imply the following well known result.

COROLLARY 3.1 ([4, 107]). *Every infinite dimensional separable Fréchet space is homeomorphic to* R^{ω}.

It is easy to observe that Polish (compact) absolute extensors are precisely retracts of R^{ω} (of I^{ω}, respectively). The following criterion helps us to recognize ANE-spaces.

THEOREM 3.4. *A space* X *is an* ANE*-space if and only if for any open cover* $\mathcal{U} \in \operatorname{cov}(X)$ *there exists a countable locally finite simplicial complex* K *(finite dimensional in case* X *is finite dimensional) and maps* $p: X \to |K|$ *and* $q: |K| \to X$ *such that the composition* qp *is* $\mathcal{U}$*-homotopic to the identity map* id_X. *In addition, we may assume that the cover* $p^{-1}(\mathcal{U}_0)$ *refines* $\mathcal{U}$, *where* $\mathcal{U}_0$ *denotes the cover of* $|K|$ *consisting of open stars (with respect to the triangulation of* K*) of vertices of* K.

Each Polish ANE-space has the homotopy type of a locally finite polyhedron [118] (the results presented in Section 3.3 imply much a stronger version of this result). Since an open subspace of an ANE-space is itself an ANE-space[9] (Proposition 2.11), we see that any open subspace of a Polish ANE-space has the homotopy type of a locally finite polyhedron. It was shown by Cauty [29] that this is in fact a characterizing property of Polish ANE-spaces.

THEOREM 3.5. *The following conditions are equivalent for any Polish space* X:

(i) X *is an* ANE*-space.*

(ii) *Each open subspace of* X *has the homotopy type of a locally finite polyhedron.*

3.1. *Examples of absolute extensors*

There are several important examples of infinite dimensional absolute extensors. Obviously, the countable powers I^{ω} and R^{ω}, of the closed and open intervals respectively, are among them. The following two spaces are also essential for further discussion:

$$\Sigma = \left\{ \{x_k\} \in l_2 \colon \sum_{k=1}^{\infty} (k x_k)^2 < \infty \right\}$$

and

$$\sigma = \big\{ \{x_k\} \in l_2 \colon \text{all but finitely many } x_k = 0 \big\}.$$

Certain topological operations preserve the property of being an absolute extensor. Some of them even improve properties of spaces in this sense.

[9] The converse of this statement is also true [100].

3.1.1. *Cones.* Recall that the cone of a space X (notation $\mathrm{Cone}(X)$) is the set $X \times [0, 1) \cup \{v\}$ topologized as follows. $X \times [0, 1)$ has the usual product topology and basic open sets containing the vertex are of the form $X \times (t, 1) \cup \{v\}$ for some $t \in [0, 1)$. In other words, $\mathrm{Cone}(X) = (X \times [0, 1]) \cup_\pi \{1\}$, where $\pi : X \times [0, 1] \to [0, 1]$ denotes the projection. $\mathrm{Cone}(X)$ is Polish for every Polish space X.

PROPOSITION 3.1. *The following conditions are equivalent*:
(i) *X is an* ANE-*space.*
(ii) $\mathrm{Cone}(X)$ *is an* AE-*space.*

THEOREM 3.6. $\mathrm{Cone}(I^\omega) \approx I^\omega$.

3.1.2. *Hyperspaces.* The hyperspace $\exp X$ of a space X is the set of all non-empty compact subsets of X endowed with the Vietoris topology. Basic open sets $\langle G_1, \ldots, G_k \rangle$ of this topology are sets of the form

$$\left\{ F \in \exp X\colon\ F \subseteq \bigcup_{i=1}^{k} G_i \text{ and } F \cap G_i \neq \emptyset \text{ for each } i = 1, \ldots, k \right\}.$$

If ρ is a metric on X, then the Hausdorff metric

$$\rho_H(F, T) = \inf\{\varepsilon > 0\colon\ F \subseteq B_\rho(T, \varepsilon) \text{ and } T \subseteq B_\rho(F, \varepsilon)\}$$

generates the same topology on $\exp X$.

Consider also the closed subspace $\exp^c X$ of $\exp X$ consisting of all subcontinua of X.

THEOREM 3.7 ([153]). *The following conditions are equivalent for any continuum X*:
(i) *X is locally connected.*
(ii) $\exp X$ *is an* AE-*compactum.*
(iii) $\exp^c X$ *is an* AE-*compactum.*

COROLLARY 3.2. *If X is an* ANE-*compactum, then* $\exp X$ *and* $\exp^c X$ *also are* ANE-*compacta.*

THEOREM 3.8 ([64]). *Let X be a nondegenerate locally connected continuum. Then*:
(i) $\exp X \approx I^\omega$.
(ii) $\exp^c X \approx I^\omega$ *if and only if X does not contain free arcs.*

THEOREM 3.9 ([63]). *The following conditions are equivalent for any space X*:
(i) $\exp X \approx R^\omega$.
(ii) *X is a connected, locally connected and nowhere locally compact Polish space.*

THEOREM 3.10 ([68]). *Let $n \geqslant 1$ and G be an abelian group. Then the subspaces* $\{F \in \exp I^\omega\colon \dim F \geqslant n\}$ *and* $\{F \in \exp I^\omega\colon \dim_G F \geqslant n\}$ *of* $\exp I^\omega$ *are homeomorphic to* Σ.

COROLLARY 3.3 ([68]). *The subspaces* $\{F \in \exp I^{\omega}\colon \dim F \geqslant \infty\}$ *and* $\{F \in \exp I^{\omega}\colon \dim_G F \geqslant \infty\}$ *of* $\exp I^{\omega}$ *are homeomorphic to* Σ^{ω}.

Let $\mathrm{ANE}(X)$ denote the subspace of $\exp X$ consisting of ANE-compacta.

THEOREM 3.11 ([69]). *If* $n \geqslant 3$, *then* $\mathrm{ANE}(R^n)$ *is homeomorphic to* Ω_3.

A related result for $n = 2$ was obtained in [31]. The space Ω_3 in the above statement is discussed in Section 3.4.1.

3.1.3. *Spaces of measures.* For a space X, let $C_B(C)$ denote the Banach space of all bounded continuous real-valued functions on X with the sup-norm and $C_B^*(X)$ be its dual with the weak*-topology. The space $C_B^*(X)$ can be identified [1, 143], by integration, with the space of signed, bounded, finitely additive, regular measures defined on the smallest algebra $\mathcal{U}(X)$ containing all closed subsets of X. Let $\Lambda(X)$ denote the set of functionals in $C_B^*(X)$ corresponding to σ-smooth measures on $\mathcal{U}(X)$. Since every σ-smooth measure on $\mathcal{U}(X)$ can be uniquely extended to a Borel measure on X (i.e., a signed, bounded, σ-additive measure defined on the smallest σ-algebra $\mathcal{B}(X)$ containing all closed subsets of X), the linear subspace $\Lambda(X) \subseteq C_B^*(X)$ can be identified with the space $M(X)$ of Borel (automatically regular in the described situation) measures on X. Further, $M^+(X)$ stands for the subspace of $M(X)$ consisting of nonnegative measures, and $P(X) = \{\mu \in M^+(X)\colon \mu(X) = 1\}$ is called the space of *probability measures*. Clearly, for a compactum X, $P(X)$ is either an n-dimensional cube (if $|X| = n + 1$) or, by [109], is homeomorphic to the Hilbert cube I^{ω} (an easier conclusion that $P(X)$ is an AE-compactum follows from Theorem 3.1). More precisely:

THEOREM 3.12 ([70, 91]). *The following conditions are equivalent for any space* X:
(i) X *is a compactum with* $|X| \geqslant \omega$.
(ii) $P(X)$ *is homeomorphic to* I^{ω}.
(iii) $M^+(X)$ *is homeomorphic to* $I^{\omega} \times [0, 1)$.

THEOREM 3.13 ([70]). *The following conditions are equivalent for any space* X:
(i) X *is a noncompact Polish space.*
(ii) $P(X)$ *is homeomorphic to* R^{ω}.
(iii) $M^+(X)$ *is homeomorphic to* R^{ω}.

One can associate with a space X various types of measures. For instance, let $M_K^+(X) \subseteq M^+(X)$ be the space of measures with compact support and $M_F^+(X) \subseteq M_K^+(X)$ be the space of measures with finite support. Also, let $P_K(X) = M_K^+(X) \cap P(X)$ and $P_F(X) = M_F^+(X) \cap P(X)$.

THEOREM 3.14 ([70]). *The following conditions are equivalent for any space* X:
(i) X *is the countable union of finite dimensional compacta.*
(ii) $P_F(X)$ *is homeomorphic to* σ.
(iii) $M_F^+(X)$ *is homeomorphic to* σ.

THEOREM 3.15 ([13,70]). *The following conditions are equivalent for any space X:*
(i) *X is noncompact, nondiscrete and locally compact.*
(ii) *$P_K(X)$ is homeomorphic to Σ.*
(iii) *$M_K^+(X)$ is homeomorphic to Σ.*

3.1.4. *Spaces of measurable functions.* For a space X a function $f:[0,1]\to X$ is said to be *measurable* if inverse images of open subsets of X are Borel subsets of $[0,1]$. If $f,g:[0,1]\to X$ are measurable, then they are equivalent if the set $\{t\in[0,1]\colon f(t)\neq g(t)\}$ has Lebesgue measure zero. The set of equivalence classes of measurable functions of $[0,1]$ into X, denoted by M_X, is topologized by the metric

$$\check{\rho}(g,g)=\sqrt{\int_0^1\big(\rho\big(f(t),g(t)\big)\big)^2\,\mathrm{d}t},$$

where ρ is a bounded metric generating the topology of X.

THEOREM 3.16 ([19]). *The following conditions are equivalent for any space X:*
(i) *X is a Polish space containing at least two points.*
(ii) *M_X is homeomorphic to R^ω.*

3.1.5. *Spaces of autohomeomorphisms.* The space of autohomeomorphisms $\mathrm{Auth}(X)$ (equipped with the compact-open topology) of a compactum X is a standard source of infinite dimensional manifolds. A comprehensive discussion of related results can be found in [96], [149]. Here we mention only few results.

THEOREM 3.17 ([113]). *Let X be a compact n-manifold, $n=1,2$. Then* $\mathrm{Auth}(X)$ *is an R^ω-manifold.*

For $n\geqslant 3$, the validity of the above statement is an open question.

THEOREM 3.18 ([93,138]). *If X is an I^ω-manifold, then* $\mathrm{Auth}(X)$ *is an R^ω-manifold.*

THEOREM 3.19 ([97,108]). *The subgroup* $\mathrm{Auth}^{PL}(X)$ *of* $\mathrm{Auth}(X)$ *consisting of PL autohomeomorphisms of a PL n-manifold X is a σ-manifold.*

3.1.6. *Spaces of continuous functions.* For a countable regular space X, let $C_p(X)$ denote the space of all continuous real-valued functions on X endowed with the topology of pointwise convergence. The topological and linear types of the spaces $C_p(X)$ have been studied extensively. It was shown in [66] that if X is nondiscrete and if $C_p(X)$ is an absolute $F_{\sigma\delta}$-set, then $C_p(X)$ is homeomorphic to $\Omega_2=\Sigma^\omega$ (see Section 3.4.1).

3.2. *Topology of I^ω-manifolds*

In this section we present all major results of I^ω-manifold theory. We begin with the characterization of the Hilbert cube manifolds which was first obtained by H. Toruńczyk.

THEOREM 3.20 (Characterization of I^ω-manifolds, [140]). *The following conditions are equivalent for a locally compact* ANE-*space* X:

(i) X *is an* I^ω-*manifold.*

(ii) X *has the* disjoint n-disk property *for each* $n \in \omega$, *i.e., the set*

$$\{f \in C(I_1^n \oplus I_2^n, X):\ f(I_1^n) \cap f(I_2^n) = \emptyset\}$$

is dense in the space $C(I_1^n \oplus I_2^n, X)$ *for each* $n \in \omega$.

(iii) *The set*

$$\{f \in C(I_1^\omega \oplus I_2^\omega, X):\ f(I_1^\omega) \cap f(I_2^\omega) = \emptyset\}$$

is dense in the space $C(I_1^\omega \oplus I_2^\omega, X)$.

(iv) *The set*

$$\{f \in C(Y, X):\ f \text{ is an embedding}\}$$

is dense in the space $C(Y, X)$ *for each compact space* Y.

(v) *The set*

$$\{f \in C(Y, X):\ f \text{ is a } Z\text{-embedding}\}$$

is dense in the space $C(Y, X)$ *for each compact space* Y.

(vi) X *is strongly universal for (locally) compact spaces (= strongly* $[\{\mathrm{pt}\}]$-*universal for (locally) compact spaces), i.e., for any (locally) compact space* B, *its closed subspace* A, *any open cover* $\mathcal{U} \in \operatorname{cov}(X)$ *and for any (proper) map* $f : B \to X$ *such that the restriction* f/A *is a* Z-*embedding, there is a* Z-*embedding* $g : B \to X$ *which is* $\mathcal{U}$-*close to* f *and such that* $g/A = f/A$.

In particular, we obtain a topological characterization of the Hilbert cube.

THEOREM 3.21 (Characterization of I^ω, [140]). *The following conditions are equivalent for a compact* AE-*space* X:

(i) X *is homeomorphic to* I^ω.

(ii) X *has the disjoint* n-*disk property for each* $n \in \omega$.

(iii) *The set*

$$\{f \in C(I_1^\omega \oplus I_2^\omega, X):\ f(I_1^\omega) \cap f(I_2^\omega) = \emptyset\}$$

is dense in the space $C(I_1^\omega \oplus I_2^\omega, X)$.

(iv) *The set*

$$\{f \in C(Y, X):\ f \text{ is an embedding}\}$$

is dense in the space $C(Y, X)$ *for each compact space* Y.

(v) *The set*

$$\{f \in C(Y, X)\colon\ f \text{ is a } Z\text{-embedding}\}$$

is dense in the space $C(Y, X)$ *for each compact space* Y.

(vi) X *is strongly universal for compact spaces, i.e., for any compact space* B, *its closed subspace* A, *any open cover* $\mathcal{U} \in \operatorname{cov}(X)$ *and for any map* $f : B \to X$ *such that the restriction* f/A *is a* Z-*embedding, there is a* Z-*embedding* $g : B \to X$ *which is* $\mathcal{U}$-*close to* f *and such that* $g/A = f/A$.

As in the case of R^{ω}-manifolds, the characterization theorem has several corollaries.

THEOREM 3.22 (ANE-theorem, [89]). *The following conditions are equivalent for any (locally) compact space* X:

(i) X *is an* A(N)E-*space.*

(ii) $X \times I^{\omega}$ *is an* I^{ω}-*manifold (respectively, is homeomorphic to* I^{ω}).

EXAMPLE 3.1. There is [117] a space X such that $X \times X \approx I^{\omega}$ but $X \not\approx I^{\omega}$.

THEOREM 3.23 ([146]). *A countable infinite product of nondegenerate* AE-*compacta is homeomorphic to* I^{ω}.

Since I^{ω} is an AE-space, the projection $X \times I^{\omega} \to X$ is a proper soft map. In particular, it is a fine homotopy equivalence (i.e., approximately soft, or, equivalently, cell-like map).

COROLLARY 3.4 ([89, 148]). *Every locally compact* ANE-*space is a proper soft image of an* I^{ω}-*manifold.*

THEOREM 3.24 (Classification by simple homotopy type, [36]). *Simply homotopy equivalent compact* I^{ω}-*manifolds are homeomorphic. Moreover, every simple homotopy equivalence of compact* I^{ω}-*manifolds is homotopic to a homeomorphism.*

REMARK 3.1. A similar statement (in terms of proper infinite simple homotopy equivalences) is true for noncompact I^{ω}-manifolds [36, 132].

THEOREM 3.25 (Classification by homotopy type, [33, 36]). I^{ω}-*manifolds* X *and* Y *are homotopy equivalent if and only if the products* $X \times [0, 1)$ *and* $Y \times [0, 1)$ *are homeomorphic. Moreover, for every homotopy equivalence* $f : X \to Y$ *the map* $f \times \mathrm{id} : X \times [0, 1) \to Y \times [0, 1)$ *is homotopic to a homeomorphism.*

COROLLARY 3.5. I^{ω} *is the only contractible compact* I^{ω}-*manifold.*

THEOREM 3.26 ([35]). *Near-homeomorphisms of* I^{ω}-*manifolds are precisely approximately soft maps.*

THEOREM 3.27 (Triangulation of I^ω-manifolds, [36, 132]). *Every I^ω-manifolds is triangulable, i.e., can be represented as the product $K \times I^\omega$ for some locally finite polyhedron K. Moreover, if an I^ω-manifold X is a Z-set of an I^ω-manifold Y, then the pair (X, Y) is triangulable, i.e., $Y = K \times I^\omega$ and $X = P \times I^\omega$ where P is a closed subpolyhedron of a polyhedron K.*

Observe that Theorems 3.22 and 3.27 imply the following result of J. West [148].

COROLLARY 3.6. *Every compact* ANE-*space is homotopy equivalent to a finite polyhedron.*

THEOREM 3.28 (Stability of I^ω-manifolds, [10]). *If X is an I^ω-manifold, then $X \times I^\omega$ is homeomorphic to X. Moreover, the projection $X \times I^\omega \to X$ is a near-homeomorphism.*

Observe that $X \approx X \times I^\omega \approx X \times I^\omega \times [0, 1] \approx X \times [0, 1]$.

THEOREM 3.29 (Open embedding theorem, [36]). *For any I^ω-manifold X the product $X \times [0, 1)$ admits an open embedding into I^ω.*

THEOREM 3.30 (Z-set unknotting, [7]). *Let X be an I^ω-manifold, A be a locally compact space and $F : A \times [0, 1] \to X$ be a proper map such that F_0 and F_1 are Z-embeddings. Then there exists an isotopy $H : X \times [0, 1] \to X$ such that $H_0 = \mathrm{id}_X$ and $H_1 F_0 = F_1$. Moreover, if F is a $\mathcal{U}$-homotopy ($\mathcal{U} \in \mathrm{cov}(X)$), then we may choose H to also be a $\mathcal{U}$-isotopy.*

Almost all results discussed above can be generalized to bundles with I^ω-fibers [142]. In particular, we have the parametric version of the characterization theorem of I^ω.

THEOREM 3.31 ([142]). *The following conditions are equivalent for a proper soft map $f : X \to Y$ between locally compact* ANE-*spaces*:
- (i) *f is a trivial bundle with fibers I^ω, i.e., f is topologically the projection $X \times I^\omega \to X$.*
- (ii) *Any proper map $g : B \to X$ of a locally compact space B into X can be approximated by Z-embeddings $h : B \to X$ with $fh = fg$.*

COROLLARY 3.7. *Let $f : X \to Y$ be a proper soft map of locally compact* ANE-*spaces. Then the composition $f\pi_X : X \times I^\omega \to X \to Y$ is a trivial bundle.*

3.3. *Topology of R^ω-manifolds*

In this section we present all major results of R^ω-manifold theory. We begin with the fundamental result of H. Toruńczyk.

THEOREM 3.32 (Characterization of R^ω-manifolds, [141]). *The following conditions are equivalent for a Polish* ANE-*space X*:

(i) *X is an R^{ω}-manifold.*
(ii) *The set*

$$\left\{f \in C\left(\bigoplus\{I^n\colon n \in \omega\}, X\right)\colon \textit{collection } \{f(I^n)\colon n \in \omega\} \textit{ is discrete in } X\right\}$$

is dense in the space $C(\bigoplus\{I^n : n \in \omega\}, X)$.
(iii) *X has the* strong discrete approximation property, *i.e., the set*

$$\left\{f \in C\left(\bigoplus\{I_n^{\omega}\colon n \in \omega\}, X\right)\colon \textit{collection } \{f(I_n^{\omega})\colon n \in \omega\} \textit{ is discrete in } X\right\}$$

is dense in the space $C(\bigoplus\{I_n^{\omega}\colon n \in \omega\}, X)$, *where* I_n^{ω} *is a copy of the Hilbert cube.*
(iv) *The set*

$$\{f \in C(Y, X)\colon f \textit{ is a closed embedding}\}$$

is dense in the space $C(Y, X)$ *for each Polish space* Y.
(v) *The set*

$$\{f \in C(Y, X)\colon f \textit{ is a } Z\textit{-embedding}\}$$

is dense in the space $C(Y, X)$ *for each Polish space* Y.
(vi) *X is strongly universal for Polish spaces (= strongly* $[\{\mathrm{pt}\}]$*-universal for Polish spaces), i.e., for any Polish space B, its closed subspace A, any open cover* $\mathcal{U} \in \mathrm{cov}(X)$ *and for any map* $f : B \to X$ *such that the restriction* f/A *is a Z-embedding, there is a Z-embedding* $g : B \to X$ *which is* $\mathcal{U}$*-close to f and such that* $g/A = f/A$.

In particular, we obtain a topological characterization of the space R^{ω}.

THEOREM 3.33 (Characterization of R^{ω}, [141]). *The following conditions are equivalent for a Polish* AE*-space X*:
(i) *X is homeomorphic to* R^{ω}.
(ii) *The set*

$$\left\{f \in C\left(\bigoplus\{I^n\colon n \in \omega\}, X\right)\colon \textit{collection } \{f(I^n)\colon n \in \omega\} \textit{ is discrete in } X\right\}$$

is dense in the space $C(\bigoplus\{I^n\colon n \in \omega\}, X)$.
(iii) *X has the strong discrete approximation property.*
(iv) *The set*

$$\{f \in C(Y, X)\colon f \textit{ is a closed embedding}\}$$

is dense in the space $C(Y, X)$ *for each Polish space* Y.

(v) *The set*

$$\{f \in C(Y, X)\colon \ f \text{ is a } Z\text{-embedding}\}$$

is dense in the space $C(Y, X)$ *for each Polish space* Y.

(vi) X *is strongly universal for Polish spaces.*

These are extremely powerful results, helpful in variety of concrete situations. Some of their immediate corollaries are very hard to prove directly. Observing that the product $X \times R^{\omega}$ always has the discrete approximation property, we get the following statement.

THEOREM 3.34 (ANE-theorem, [137]). *The following conditions are equivalent for any Polish space* X:

(i) X *is an* A(N)E-*space.*

(ii) $X \times R^{\omega}$ *is an* R^{ω}*-manifold (respectively, is homeomorphic to* R^{ω}*).*

EXAMPLE 3.2 ([8]). There is a space X such that $X \times X \approx R^{\omega}$ but $X \not\approx R^{\omega}$.

THEOREM 3.35 ([141]). *A countable infinite product of noncompact Polish* AE-*spaces is homeomorphic to* R^{ω}.

Since R^{ω} is an AE-space, the projection $X \times R^{\omega} \to X$ is a soft map. In particular, it is a fine homotopy equivalence (i.e., approximately soft).

THEOREM 3.36. *Every Polish* ANE-*space is a soft image of an* R^{ω}*-manifold.*

THEOREM 3.37 (Homotopy classification, [106]). *Homotopy equivalent* R^{ω}*-manifolds are homeomorphic. Moreover, every homotopy equivalence between* R^{ω}*-manifolds is homotopic to a homeomorphism.*

COROLLARY 3.8. R^{ω} *is the only contractible* R^{ω}*-manifold.*

THEOREM 3.38 ([141]). *Near-homeomorphisms of* R^{ω}*-manifolds are precisely approximately soft maps.*

COROLLARY 3.9 (Stability of R^{ω}-manifolds, [10, 129]). *If* X *is an* R^{ω}*-manifold, then* $X \times R^{\omega}$ *is homeomorphic to* X. *Moreover, the projection* $X \times R^{\omega} \to X$ *is a near-homeomorphism.*

Observe that $X \approx X \times R^{\omega} \approx X \times R^{\omega} \times [0, 1] \approx X \times [0, 1]$.

COROLLARY 3.10. *If* A *is a* Z_{σ}*-set in an* R^{ω}*-manifold* X, *then* $X - A$ *is an* R^{ω}*-manifold. Moreover, the inclusion map* $X - A \hookrightarrow X$ *is a near-homeomorphism.*

THEOREM 3.39 (Triangulation of R^ω-manifolds, [105]). *Every R^ω-manifold is triangulable, i.e., can be represented as the product $K \times R^\omega$ for some locally finite polyhedron K. Moreover, if an R^ω-manifold X is a Z-set of an R^ω-manifold Y, then the pair (X, Y) is triangulable, i.e., $Y = K \times R^\omega$ and $X = P \times R^\omega$ where P is a closed subpolyhedron of a polyhedron K.*

THEOREM 3.40 (Open embedding theorem, [104, 105]). *R^ω-manifolds are precisely open subspaces of R^ω.*

THEOREM 3.41 (Z-set unknotting, [9]). *Let X be an R^ω-manifold, A be a Polish space, and $F : A \times [0, 1] \to X$ be a map such that F_0 and F_1 are Z-embeddings. Then there exists an isotopy $H : X \times [0, 1] \to X$ such that $H_0 = \mathrm{id}_X$ and $H_1 F_0 = F_1$. Moreover, if F is an $\mathcal{U}$-homotopy ($\mathcal{U} \in \mathrm{cov}(X)$), then we may choose H to also be a $\mathcal{U}$-isotopy.*

It is possible to characterize trivial bundles with fibers R^ω.

THEOREM 3.42 (Toruńczyk, West). *The following conditions are equivalent for a soft map $f : X \to Y$ between Polish* ANE-*spaces*:

(i) *f is a trivial bundle with fibers R^ω, i.e., f is topologically the projection $X \times R^\omega \to X$.*
(ii) *Any map $g : B \to X$ of a Polish space B into X can be approximated by Z-embeddings $h : B \to X$ with $fh = fg$.*

COROLLARY 3.11. *Let $f : X \to Y$ be a soft map of Polish* ANE-*spaces. Then the composition $f\pi_X : X \times R^\omega \to X \to Y$ is a trivial bundle.*

3.4. *Topology of Σ-manifolds*

In this section we present results about Σ- and σ-manifolds. In many respect they are similar to those from Sections 3.3 and 3.2.

THEOREM 3.43 (Characterization of Σ-manifolds, [22, 120]). *An* ANE-*space X is a Σ-manifold if and only if the following conditions are satisfied*:

(i) *X is a countable union of compacta.*
(ii) *X is a countable union of strong Z-sets.*
(iii) *X is strongly universal for the class of countable unions of compacta, i.e., for any space B which can be written as a countable union of compacta, its closed subspace A, any open cover $\mathcal{U} \in \mathrm{cov}(X)$ and any map $f : B \to X$, such that the restriction f/A is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.*

In particular, we obtain a topological characterization of the space Σ.

THEOREM 3.44 (Characterization of Σ, [22, 120]). *An* AE-*space X is homeomorphic to Σ if and only if the following conditions are satisfied*:

(i) *X is a countable union of compacta.*

(ii) *X is a countable union of strong Z-sets.*

(iii) *X is strongly universal for the class of countable unions of compacta, i.e., for any space B which can be written as a countable union of compacta, its closed subspace A, any open cover $\mathcal{U} \in \mathrm{cov}(X)$ and any map $f : B \to X$, such that the restriction f/A is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.*

We list some of the corollaries of the characterization theorems.

THEOREM 3.45 (ANE-theorem, [89]). *The following conditions are equivalent for any space X:*

(i) *X is an* A(N)E*-space which is a countable union of compacta.*

(ii) *$X \times \Sigma$ is a Σ-manifold (respectively, is homeomorphic to Σ).*

THEOREM 3.46 (Classification by homotopy type, [32]). *Homotopy equivalent Σ-manifolds are homeomorphic.*

COROLLARY 3.12. *Σ is the only contractible compact Σ-manifold.*

THEOREM 3.47. *Near-homeomorphisms of Σ-manifolds are precisely approximately soft maps.*

THEOREM 3.48 (Triangulation of Σ-manifolds, [32]). *Every Σ-manifold is triangulable, i.e., can be represented as the product $K \times \Sigma$ for some locally finite polyhedron K.*

THEOREM 3.49 (Stability of Σ-manifolds, [32]). *If X is a Σ-manifold, then X, $X \times I^n$, $X \times I^\omega$ and $X \times \Sigma$ are homeomorphic.*

THEOREM 3.50 (Open embedding theorem, [32]). *The class of Σ-manifolds coincides with the class of open subspaces of Σ.*

THEOREM 3.51 (Z-set unknotting, [32]). *Let X be a Σ-manifold. Then for each open cover $\mathcal{U} \in \mathrm{cov}(X)$ there is an open cover $\mathcal{V} \in \mathrm{cov}(X)$ such that the following condition is satisfied:*

(*) *For any homeomorphism $h : A \to B$ between Z-sets of X which is $\mathcal{V}$-close to the inclusion map $A \hookrightarrow X$ there is a homeomorphism $H : X \to X$ which is $\mathcal{U}$-close to id_X and such that $H/A = h$.*

By replacing in the results listed above arbitrary compacta by finite dimensional ones we obtain similar statements for the space σ.

3.4.1. *Other incomplete manifolds.* The spaces I^ω, R^ω and Σ are representatives of the first three classes of the hierarchy of *Borel classes*. For a countable ordinal α there exist the absolute additive Borelian class $\mathcal{A}_\alpha$ and the absolute multiplicative Borelian class $\mathcal{M}_\alpha$.

Note that $\mathcal{A}_0 = \emptyset$; $\mathcal{M}_0$ consists of all compacta; $\mathcal{A}_1$ of all σ-compact spaces; $\mathcal{M}_1$ is the collection of all Polish spaces, etc. The concept of strong universality which was exploited in Theorems 3.20, 3.32 and 3.43 can be naturally adapted to each of the above classes. It was shown in [22] that for each $\alpha \geqslant 1$, there exists an AE-space Ω_α which is strongly universal with respect to spaces from the class $\mathcal{M}_\alpha$. Similarly, for each $\alpha \geqslant 1$ there exists an AE-space Λ_α which is strongly universal with respect to spaces from the class $\mathcal{A}_\alpha$ (in this notation $\Sigma = \Lambda_1$). These spaces can be characterized topologically as follows.

THEOREM 3.52 ([22]). *Let $\alpha \geqslant 1$ and X be an* AE-*space. Then the following conditions are equivalent*:

(i) *X is homeomorphic to Ω_α (respectively, $X \approx \Lambda_\alpha$).*
(ii) *$X = \bigcup_{i=1}^{\infty} X_i$, where $X_i \in \mathcal{M}_\alpha$ (respectively, $X_i \in \mathcal{A}_\alpha$) and X_i is a strong Z-set in X for each i.*
(iii) *X is strongly universal with respect to spaces from the class $\mathcal{M}_\alpha$ (respectively, from the class $\mathcal{A}_\alpha$).*

3.5. *Shape theory*

The shape category SHAPE was defined in Section 2.7 (take $L = \{\mathrm{pt}\}$). There are several important concepts in shape theory (such as strong shapes, algebraic invariants of shape theory, etc.) which are not even discussed here. We refer to [25, 84, 85, 115] and [149] where the reader can find detailed discussions of various topics and many open problems. Here we concentrate on the geometric side of this concept. Chapman's famous result gives its geometric interpretation.

THEOREM 3.53 (Complement theorem, [33]). *Let X and Y be Z-sets in the Hilbert cube I^ω. Then the following conditions are equivalent*:

(i) *X and Y have the same shape.*
(ii) *The complements $I^\omega - X$ and $I^\omega - Y$ are homeomorphic.*
(iii) *The complements $I^\omega - X$ and $I^\omega - Y$ are weakly properly homotopy equivalent.*[10]

The equivalence of conditions (ii) and (iii), in comparison with Remark 3.1 and Theorem 3.25, shows that complements of Z-sets in I^ω are I^ω-manifolds of very special type. In order to obtain their internal characterization, let us say that an I^ω-manifold X *has a boundary* if there exists a compact I^ω-manifold $\widetilde{X}$ such that $X \subseteq \widetilde{X}$ and the complement $\widetilde{X} - X$ is a Z-set in $\widetilde{X}$. In this situation we say that $\widetilde{X} - X$ is a *boundary* of X in $\widetilde{X}$. The problem of characterizing of I^ω-manifolds with boundaries has its roots in earlier works [28, 131] on a similar problem for usual manifolds. For any contractible open subspace W of R^3 not homeomorphic to R^3 the product $W \times I^\omega$ produces an example of a contractible I^ω-manifold without boundary. It turns out [39] that if an I^ω-manifold X satisfies certain necessary homotopy theoretic conditions (finite type and tamenes at ∞), then there exists an obstruction $\beta(X)$ to X having a boundary. $\beta(X)$ lies in the quotient

[10] See p. 28.

$S_\infty(X)$ of the group $S(X)$ of all infinite simple homotopy types on X by the image of the Whitehead group $Wh\pi_1(X)$. This group depends only on the inverse system

$$\{\pi_1(X-K)\colon\ K \text{ is a compact subset of } X\}.$$

It turns out that the different boundaries that can be put on X constitute a whole shape class. The different ways of putting boundaries on X can be classified by elements of the inverse limit of the above indicated inverse system.

The next step in this direction is to ask when I^ω-manifolds (with boundaries) admit a polyhedral boundary? This problem was considered and solved in [38]. There are additional obstructions to an I^ω-manifold having a polyhedral boundary. This obstruction can be understood via the following result.

THEOREM 3.54 (Proper finiteness obstruction theorem, [38]). *If X is an I^ω-manifold which is finitely dominated near ∞, then there is an element $\sigma_p(X)$ in*[11]

$$\widetilde{K}_0\pi_1 E(X) = \varprojlim\{\widetilde{K}_0\pi_1(X-K)\colon\ K \textit{ is a compact subset of } X\}$$

which vanishes if and only if X has finite type near ∞. Moreover, $\sigma_p(X)$ is an invariant of proper homotopy type near ∞.

Several homotopy theoretical questions for complexes have their counterparts in shape theory. For instance, if a continuum X is shape dominated by a finite complex, is it shape equivalent to a finite complex? As in the case of homotopies, the answer is not always affirmative [90]. There is an intrinsically defined shape theoretic version of Wall's obstruction $w(X) \in \widetilde{K}_0\check{\pi}_1(X)$ which vanishes if and only if X is shape equivalent to a finite complex (here $\check{\pi}_1(X)$ denotes the Čech fundamental group of X).

4. The case $L = S^n$: "n-dimensional" infinite dimensional topology

In this section we discuss the case $L = S^n$ which, as was mentioned above, produces the theory of n-dimensional Menger and Nöbeling manifolds. Clearly ANE($[S^n]$)-spaces are precisely LC^{n-1}-spaces, $UV^{[S^n]}$-maps are called UV^n-maps (not UV^{n-1}-maps as usual), etc. We recall that $[S^n]$-homotopies and $[S^n]$-shapes in this paper are called n-homotopies and n-shapes, respectively (instead of the originally used terminology: $(n-1)$-homotopies and $(n-1)$-shapes). Of course, the Menger compactum $\mu^{[S^n]}$ modulo $[S^n]$ is denoted, as usual, by μ^n. The same applies to the Nöbeling space modulo $[S^n]$: $\nu^n \equiv \nu^{[S^n]}$.

4.1. *Menger manifolds*

We begin with the standard construction of the Menger compactum. Let $n \geqslant 0$. We partition the unit cube I^{2n+1}, lying in $(2n+1)$-dimensional Euclidean space R^{2n+1}, into 3^{2n+1}

[11] Recall that Wall's finiteness obstruction is an element of the reduced projective class group $\widetilde{K}_0\pi_1(X)$.

congruent cubes of the "first rank" by hyperplanes drawn perpendicular to the edges of the cube I^{2n+1} at points dividing the edges into three equal parts, and we choose from these 3^{2n+1} cubes those which intersect the n-dimensional skeleton of the cube I^{2n+1}. The union of the selected cubes of the "first rank" is denoted by $I(n,1)$. In an analogous way, we divide every cube entering as a term in $I(n,1)$ into 3^{2n+1} congruent cubes of the "second rank", and the union of all analogously selected cubes of the "second rank" is denoted by $I(n,2)$. If we continue the process we get a decreasing sequence of compacta

$$I(n,1) \supseteq I(n,2) \supseteq \cdots.$$

The compactum $\mu^n = \bigcap\{I(n,i)\colon n = 1,2,\ldots\}$ is called the *n-dimensional universal Menger compactum*. Note that $\mu^n = M_n^{2n+1}$, where M_n^k, $0 \leqslant n < k$, denotes the "n-dimensional Menger compactum constructed in the k-dimensional cube I^k" (a precise definition of M_n^k is given below).

Lefschetz's construction [112] of Menger compacta is sometimes more convenient.

Let M be a PL k-manifold with a (combinatorial) triangulation L. Inductively, we define a sequence $\{M_i\}$ of PL k-manifolds and their triangulations L_i as follows.

Let $M_0 = M$ and $L_0 = L$. Let $M_1 = \mathrm{st}(L^{(n)}, \beta^2 L_0)$, $L_1 = \beta^2 L_0|M_1$, and suppose that M_i and L_i have already been defined. Consider $\beta^2 L_i$, and let $M_{i+1} = \mathrm{st}(L_i^{(n)}, \beta^2 L_i)$ and $L_{i+1} = \beta^2 L_i|M_{i+1}$. Then $\{M_i\}$ is a decreasing sequence and $\bigcap M_i \neq \emptyset$. If M is the k-simplex with the standard simplicial complex structure, then the resulting compactum $\bigcap M_i$ is denoted by L_n^k. Notice that M_{i+1} may be regarded as a regular neighborhood of the n-skeleton of M_i (with respect to L_i).

Obviously, μ^0 coincides with the Cantor discontinuum and, consequently, is the only zero-dimensional compactum with no isolated points.

Positive dimensional Menger compacta M_n^k were originally defined within the classical dimension theory by Menger in [116]. They are generalizations of the Cantor discontinuum and Sierpinski's universal curve M_1^2 [133]. Recall that the space M_1^2 is universal for the class of all at most 1-dimensional planar compacta [133]. Further, it was shown by Menger [116] that the 1-dimensional Menger compactum $\mu^1 = M_1^3$ is universal for the class of all at most 1-dimensional compacta. Generally, it was conjectured [116] that M_n^k is a universal space for the class of all at most n-dimensional compacta embeddable in R^k (Menger's problem). As was already mentioned, this problem was known to have a positive solution for $(n,k) = (1,2)$ and $(n,k) = (1,3)$. A positive solution in the case $n = k-1$ was also given by Menger [116]. Results of Lefschetz [112] and Bothe [27] produced a positive solution in the case $k = 2n+1$. The ultimate affirmative solution of Menger's problem was obtained by Štanko [135].

The 1-dimensional Menger compacta M_1^2 and $\mu^1 = M_1^3$ were also characterized topologically by Whyburn [152] and Anderson [2] (see also [3]), respectively. Anderson's theorem characterizes μ^1 as a 1-dimensional locally connected continuum with no local separating points and with no nonempty open subspaces embeddable into the plane.

To the best of my knowledge, there are no reasonable conjectures concerning the characteristic properties of the compacta M_n^k when $1 < n < k \leqslant 2n$.

THEOREM 4.1 (Characterization of μ^n-manifolds, [20]). *The following conditions are equivalent for any n-dimensional locally compact* ANE(n)*-space X*:

(i) *X is a μ^n-manifold.*

(ii) *X has the disjoint n-disk property.*

(iii) *The set*

$$\{f \in C(Y, X)\colon\ f \text{ is an embedding}\}$$

is dense in the space $C(Y, X)$ for each at most n-dimensional compact space Y.

(iii) *The set*

$$\{f \in C(Y, X)\colon\ f \text{ is a } Z\text{-embedding}\}$$

is dense in the space $C(Y, X)$ for each at most n-dimensional compact space Y.

(iv) *X is strongly universal for at most n-dimensional (locally) compact spaces (= strongly $[S^n]$-universal for (locally) compact spaces), i.e., for any at most n-dimensional (locally) compact space B, its closed subspace A, any open cover $\mathcal{U} \in \operatorname{cov}(X)$ and any (proper) map $f : B \to X$ such that the restriction f/A is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.*

Additionally, if X is compact and $(n-1)$-connected (i.e., $X \in \mathrm{AE}([S^n]) = \mathrm{AE}(n)$), then conditions (ii)–(vi) give a topological characterization of the compactum μ^n.

THEOREM 4.2 (Resolution theorem, [20,72,73]). *There are maps $f_n : \mu^n \to I^\omega$ and $g_n : \mu^n \to \mu^n$ satisfying the following properties*:

(i) *The maps f_n and g_n are n-invertible, $(n-1)$-soft and polyhedrally n-soft.*

(ii) *All the fibers of the maps f_n and g_n are homeomorphic to μ^n.*

(iii) *The inverse images of* ANE(n)*-compacta under the maps f_n and g_n are μ^n-manifolds.*

(iii) *The maps f_n and g_n both satisfy the parametric version of the disjoint n-disk property, that is, any two maps $\alpha, \beta : I^n \to \mu^n$ can be arbitrarily closely approximated by maps $\alpha', \beta' : I^n \to \mu^n$ such that $f_n\alpha' = f_n\alpha$, $f_n\beta' = f_n\beta$, $g_n\alpha' = g_n\alpha$, $g_n\beta' = g_n\beta$ and $\operatorname{im}(\alpha') \cap \operatorname{im}(\beta') = \emptyset$.*

At first glance it is not clear what is the analog of the operation of "taking the product by I^ω" in μ^n-manifold theory – the operation which is involved in the formulations of the triangulation (Theorem 3.27) and stability (Theorem 3.28) theorems for I^ω-manifolds. Taking the product $X \times I^\omega$ of a space X and the Hilbert cube I^ω may be interpreted as taking the inverse image $\pi_1^{-1}(X)$ of a space $X \subseteq I^\omega$, where $\pi_1 : I^\omega \times I^\omega \to I^\omega$ denotes the natural projection onto the first coordinate. It turns out that the map $g_n = f_n/f_n^{-1}(\mu^n) : \mu^n \to \mu^n$ in Theorem 4.2 may be thought of as the analog of the projection π_1 in the theory of μ^n-manifolds. If this is agreed, everything then falls in place.

THEOREM 4.3 (Triangulation of μ^n-manifolds, [73]). *For any μ^n-manifold M, there is an n-dimensional polyhedron P such that for any embedding of P into μ^n the inverse image $g_n^{-1}(P)$ is homeomorphic to M.*

THEOREM 4.4 (Stability of μ^n-manifolds, [73]). *For any μ^n-manifold M in μ^n, the inverse image $g_n^{-1}(M)$ is homeomorphic to M.*

THEOREM 4.5 (Classification by proper n-homotopy type, [20]). *Properly n-homotopic μ^n-manifolds are homeomorphic. Moreover, every proper n-homotopy equivalence of μ^n-manifolds is properly n-homotopic to a homeomorphism.*

THEOREM 4.6 (Classification by n-homotopy type, [48]). *Two μ^n-manifolds are n-homotopy equivalent if and only if their n-homotopy kernels*[12] *are homeomorphic.*

THEOREM 4.7 ([20]). *Near-homeomorphisms of μ^n-manifolds are precisely proper approximately n-soft maps ($=$ proper UV^n-maps).*

THEOREM 4.8 (Open embedding theorem, [48]). *The n-homotopy kernel of an μ^n-manifold admits an open embedding into μ^n.*

THEOREM 4.9 (Local Z-set unknotting, [20]). *Let X be a μ^n-manifold. Then for each open cover $\mathcal{U} \in \mathrm{cov}(X)$ there is an open cover $\mathcal{V} \in \mathrm{cov}(X)$ such that the following condition is satisfied:*

(∗) *For any homeomorphism $h : A \to B$ of Z-sets of X such that h is $\mathcal{V}$-close to the inclusion map $A \hookrightarrow X$, there is a homeomorphism $H : X \to X$ which is $\mathcal{U}$-close to id_X and such that $H/A = h$.*

THEOREM 4.10 (Global Z-set unknotting, [20]). *Let X be a μ^n-manifold and let $h : A \to B$ be a homeomorphism between Z-sets of X which is properly n-homotopic to the inclusion map $A \hookrightarrow X$. Then there exists a homeomorphism $H : X \to X$ which is properly n-homotopic to id_X and such that $h = H/A$.*

COROLLARY 4.1 (Homogeneity of μ^n, [2,3,20]). *The universal Menger compactum μ^n is topologically homogeneous.*

4.2. *Nöbeling manifolds*

Much less is known about the n-dimensional Nöbeling space ν^n. Let us first recall the definition of the space N_n^k ($0 \leqslant n < k \leqslant \infty$). This is the subspace of k-dimensional Euclidean space R^k consisting of points at most n coordinates of which are rational. In what follows, we consider only the cases $0 < 2n+1 \leqslant k \leqslant \infty$ and use the notation $\nu^n = N_n^{2n+1}$. We repeat here Conjecture 2.6 (with $L = S^n$) which was asked by several authors [51,67,149].

[12] See p. 337 for definition of $[L]$-homotopy kernel.

CONJECTURE 4.1. A Polish space X is homeomorphic to ν^n provided that X satisfies the following conditions:

(i) $\dim X = n$.

(ii) $X \in \mathrm{AE}([S^n]) = \mathrm{AE}(n) = LC^{n-1} \cap C^{n-1}$.

(iii) X is strongly universal for at most n-dimensional Polish spaces (= strongly $[S^n]$-universal for Polish spaces), i.e., for any at most n-dimensional Polish space B, its closed subspace A, any open cover $\mathcal{U} \in \mathrm{cov}(X)$ and for any map $f : B \to X$ such that the restriction f/A is a Z-embedding, there is a Z-embedding $g : B \to X$ which is $\mathcal{U}$-close to f and such that $g/A = f/A$.

It was observed in [136] that $R^{2n+1} - \nu^n$ is a skeletoid with respect to a certain collection of n-dimensional polyhedra in R^{2n+1} (see [98] for the construction of different skeletoids in R^k). The compactum μ^n also contains a skeletoid Σ^n, which is nothing else but the space $\Sigma^{[S^n]}$ constructed on p. 33.

THEOREM 4.11 ([59]). *ν^n is homeomorphic to the pseudo-interior $\mu^n - \Sigma^n$ of the Menger compactum μ^n.*

Most of the results listed below are based on the above theorem.

COROLLARY 4.2 ([59,65]). *Let $k \geqslant 2n + 1$. Then ν^n is homeomorphic to the pseudo-interior s_n^k of R^k.*

COROLLARY 4.3. *The space ν^n is topologically homogeneous.*

According to [51], every μ^n-manifold M contains a skeletoid $\Sigma^n(M)$. Their complements $\nu^n(M) = M - \Sigma^n(M)$ are ν^n-manifolds of a special type. Let us call them *regular ν^n-manifolds*. It is not known whether every ν^n-manifold is regular.

THEOREM 4.12 ([51]). *The following conditions are equivalent for any space X:*

(i) *X is homeomorphic to an open subspace of ν^n.*

(ii) *X is a regular ν^n-manifold.*

THEOREM 4.13 (Classification by n-homotopy type, [51]). *n-homotopy equivalent regular ν^n-manifolds are homeomorphic.*

The following corollary of the above theorem deserves to be stated separately (I have no idea how to prove it directly).

COROLLARY 4.4 ([51]). *An $(n - 1)$-connected open subspace of ν^n is homeomorphic to ν^n. In particular, a connected open subspace of ν^1 is homeomorphic to ν^1.*

4.3. *n-shape theory*

The n-shape category n-SHAPE can be defined by following the general definition (Section 2.7, the case $L = S^n$). Originally this concept was defined in [43] (compare with [94]). Isomorphic objects of this category are said to have the same n-shape.

THEOREM 4.14 (Complement theorem for μ^n, [43]). *Let X and Y be Z-sets in the Menger compactum μ^n. Then the following conditions are equivalent*:

(i) *The complements $\mu^n - X$ and $\mu^n - Y$ are homeomorphic.*

(ii) $n - \mathrm{Sh}(X) = n - \mathrm{Sh}(Y)$, *i.e., X and Y have the same n-shape.*

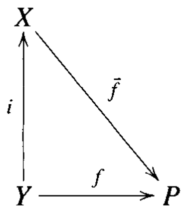
All results discussed in Section 3.5 have their natural counterparts in the n-shape theory. We refer the reader to [57] for details. Here we only remark that the natural analogs of algebraic obstructions (see Section 3.5) corresponding to the n-shape category vanish. The explanation of this phenomena is provided in the next statement (see also p. 335) which shows that the analog of Wall's obstruction vanishes in the n-homotopy category.

THEOREM 4.15 ([50]). *If a locally finite polyhedron is n-homotopy dominated by an at most n-dimensional finite polyhedron, then it is n-homotopy equivalent to an n-dimensional finite polyhedron.*

5. Non-metrizable manifolds

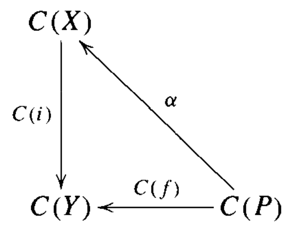
Non-metrizable manifolds are traditionally outside of what is called Geometric Topology. On the other hand, the uncountable powers I^τ and R^τ of the closed segment and of the real line are extremely important and interesting objects from various points of view and deserve to be discussed here. Below we present a brief outline of an extension of the classical ANE-theory (case $L = \{\mathrm{pt}\}$) to the nonmetrizable case. One of the main differences of a technical nature between the metrizable and nonmetrizable parts of the ANE-theory is that the apparatus of uncountable inverse spectra becomes the main tool of investigation in the nonmetrizable case. We recommend [51] for a comprehensive discussion of related topics. There is another, but not less essential, difference hidden in the classical definition of ANE-spaces. Indeed, according to the definition of the relation $X \in \mathrm{AE}(Y)$ (see Section 2.1) the real line is not an AE-space in the class of Tychonov (i.e., completely regular and Hausdorff) spaces. This fact immediately follows from the simple observation that a nonnormal space always contains a closed subspace and a continuous real-valued function not extendable onto the whole space. In order to modify this definition, let us analyze closely the classical extension problem. The most general form of this problem asks:

EP: Does a map $f : Y \to P$, defined on a (closed) subspace Y of a space X, admit an extension $\bar{f} : X \to P$



Consider the category TYCH of Tychonov spaces and continuous maps on the one hand, and the category VECT of vector spaces and linear maps on the other. There is a natural contravariant functor $C: \mathrm{TYCH} \to \mathrm{VECT}$ assigning to each space X the vector space $C(X)$ of all continuous real-valued functions on X, and to each continuous map $f: X \to Y$ the induced linear operator $C(f): C(Y) \to C(X)$. Applying the functor C to the above Extension Problem (see [134, Theorem I.1.10]) we arrive at the corresponding Lifting Problem in the category VECT:

LP: Does the map $C(f): C(P) \to C(Y)$ have a lifting $\alpha: C(P) \to C(X)$ (i.e., does there exist a map α such that $C(f) = C(i)\alpha$)?



The existence of such a lifting produces the following necessary condition for EP to have a solution.

NC: $C(f)(C(P)) = C(i)(\alpha(C(P)) \subseteq C(i)(C(X)) = C(X)/Y$, where $C(X)/Y = \{\varphi \in C(Y)$: there exists $\bar{\varphi} \in C(X)$ such that $\bar{\varphi}/Y = \varphi\}$.

Obviously, by the Tietze–Urysohn Extension Theorem, condition NC is automatically satisfied for any $f: Y \to P$ defined on a closed subspace Y of a normal space X.

Now recall that a subset A of a space X is called *functionally open* (*closed*) if there is a function $f: X \to R$ such that $A = f^{-1}(R - \{0\})$ (respectively, $A = f^{-1}(0)$). Let us call a functionally open neighborhood U of a subset A in a space X *stable* if there is a functionally closed subset Z of X such that $A \subseteq Z \subseteq U$. Observe that if a subset A is C-embedded in X, then every functionally open neighborhood of A is stable.

Based on the above observation we arrive at the following definition.

DEFINITION 5.1. Let X and L be Tychonov spaces. We say that L is an absolute (neighborhood) extensor of X, and write $L \in \mathrm{A(N)E}(X)$, if for each subspace Y of X any map $f: Y \to L$ such that $C(f)(C(L)) \subseteq C(X)/Y$ admits an extension $\tilde{f}: X \to L$ (respectively, $\tilde{f}: G \to L$) over X (respectively, over a stable neighborhood G of A in X).

Note that we have dropped the requirement that Y be closed in X.

This approach allows us to define the extension dimension of any space[13] and to extend the theory of ANE($[L]$)-spaces to the nonmetrizable case. Realcompact spaces play an important role in this theory. These are spaces homeomorphic to closed subspaces of powers of the real line. Information about realcompact spaces, and in particular about Hewitt realcompactification υX, can be found in [51, 99].

THEOREM 5.1 ([53]). *Let X be a (Tychonov) space and L be a countable complex. Then $L \in \mathrm{AE}(X)$ if and only if $L \in \mathrm{AE}(\upsilon X)$.*

THEOREM 5.2 ([53]). *The following conditions are equivalent for any realcompact space X and a countable complex L:*

(i) $\mathrm{ed}(X) \leqslant [L]$.
(ii) *X is the limit space of a Polish spectrum $\mathcal{S}_X = \{X_\alpha, p_\alpha^\beta, A\}$ such that $\mathrm{ed}(X_\alpha) \leqslant [L]$ for each $\alpha \in A$.*

Theorems 5.1 and 5.2 reduce the theory of extension dimension of general spaces to that of Polish spaces.

Definition 2.2 can also be extended to the nonmetrizable case.

DEFINITION 5.2. Let L be a complex. A space X is called an absolute (neighborhood) extensor modulo $[L]$ (shortly, A(N)E($[L]$)-space) if for each space Z with $\mathrm{ed}(Z) \leqslant [L]$ and each subspace Z_0 of Z, any map $f : Z_0 \to X$, such that $C(f)(C(X)) \subseteq C(Z)/Z_0$, can be extended to (some stable functionally open neighborhood of Z_0) Z.

Since $[L] \geqslant [S^0]$, every ANE($[L]$)-space is an AE($[S^0]$)-space (i.e., AE(0)-space in the terminology of [51]). Therefore ANE($[L]$)-spaces have all the properties of AE(0)-spaces discussed in [51]. In particular, such spaces are realcompact. Within the class of realcompact spaces, Definition 2.3 admits the following extension.

DEFINITION 5.3. Let L be a complex. A map $f : X \to Y$ of a realcompact spaces is said to be $[L]$-soft if for each realcompact space Z with $\mathrm{ed}(Z) \leqslant [L]$, for its (closed) subspace Z_0, and for any two maps $g : Z_0 \to X$ and $h : Z \to Y$ such that $fg = h/Z_0$ and $C(g)(C(X)) \subseteq C(Z)/Z_0$, there exists a map $k : Z \to X$ such that $k/Z_0 = g$ and $fk = h$.

Obviously the class of AE($[L]$)-spaces coincides with the class of realcompact spaces, the constant maps of which are $[L]$-soft. Using techniques of [51, 53] (see also related papers [41, 42]) one can prove the following statements.

PROPOSITION 5.1. *Let L be a complex and τ be an infinite cardinal. Then there is an $[L]$-soft map $f_{[L],\tau} : X_{[L],\tau} \to R^\tau$ satisfying the following conditions:*

(i) $\mathrm{ed}(X_{[L],\tau}) = [L]$.
(ii) $w(X_{[L],\tau}) = \tau$.

---
[13] In this section all spaces are Tychonov.

The above proposition is a key ingredient of the proof of the following statement which characterizes nonmetrizable A(N)E($[L]$)-spaces.

THEOREM 5.3. *Let L be a complex. A space X of weight $\tau \geqslant \omega$ is an* A(N)E($[L]$)*-space if and only if it is the limit space of an inverse spectrum $\mathcal{S}_X = \{X_\alpha, p_\alpha^{\alpha+1}, \tau\}$ satisfying the following conditions*:

(i) *X_0 is a Polish* A(N)E($[L]$)*-space.*
(ii) *Each short projection $p_\alpha^{\alpha+1} : X_{\alpha+1} \to X_\alpha$ is $[L]$-soft.*
(iii) *Each short projection $p_\alpha^{\alpha+1}$ has a Polish kernel in the sense of* [51].
(iv) *The spectrum $\mathcal{S}_X$ is continuous, i.e., for each limit ordinal $\alpha < \tau$, the space X_α is canonically homeomorphic to the limit space of the spectrum $\{X_\beta, p_\beta^{\beta+1}, \alpha\}$.*

For $[L] = [S^0]$ and compact X, the above result was proved in [103]. For $[L] = [S^n]$ (with $n \geqslant 1$) and compact X, it was obtained in [72] (see also [92] and [121]). The case $[L] = [\{\text{pt}\}]$ and compact X was considered in [126]. All these particular results were generalized for noncompact spaces in [41]. Theorem 5.3 has various applications. Proofs of most of the results below are based on this statement.

Here is another, sometimes more convenient, version of the above statement.

THEOREM 5.4. *Let L be a complex and $\tau \geqslant \omega$. A space X is an* A(N)E($[L]$)*-space if and only if it is the limit space of a τ-spectrum $\mathcal{S}_X = \{X_\alpha, p_\alpha^\beta, \tau\}$ satisfying the following conditions*:

(i) *X_α is an* A(N)E($[L]$)*-space of weight τ, $\alpha \in A$.*
(ii) *Each limit projection $p_\alpha : X \to X_\alpha$ is $[L]$-soft.*

Theorem 5.3 and Proposition 2.17 imply the following result.

THEOREM 5.5 ([61]). *Let L be a noncontractible connected complex and X be an* ANE($[L]$)*-compactum. If* $\text{ed}(X) \leqslant [L]$, *then X is metrizable.*

The case $[L] = [S^n]$ was earlier proved by Dranishnikov [72]. Theorem 5.5 strengthens also the result of Ščepin [127] on metrizability of finite-dimensional ANE-compacta.

Note that unlike the compact case, there exist noncompact AE($[L]$)-spaces of extension dimension $[L]$ and of arbitrary weight.

5.1. *τ-limitation topology*

Let X and Y be arbitrary Tychonov spaces and τ be an arbitrary infinite cardinal number. We are going to introduce a topology (τ-limitation topology, depending on τ) on the set $C(Y, X)$ of all continuous maps from Y into X. The space thus obtained will be denoted by $C_\tau(Y, X)$. We recall that $\text{cov}(X)$ denotes the collection of all countable functionally open covers of the space X. For each map $f : Y \to X$ the sets of the form

$$B\big(f, \{\mathcal{U}_t\colon\ t \in T\}\big) = \big\{g \in C(Y, X)\colon\ g \text{ is } \mathcal{U}_t\text{-close to } f \text{ for each } t \in T\big\},$$

where $|T| < \tau$ and $\mathcal{U}_t \in \text{cov}(X)$ for each $t \in T$, are declared to be open basic neighborhoods of the point f in $C_\tau(Y, X)$.

The maps contained in the neighborhood $B(f, \{\mathcal{U}_t\colon t \in T\})$ are called $\{\mathcal{U}_t\colon t \in T\}$-close to f.

If a space X has a countable basis, then, obviously, the space $C_\omega(Y, X)$ (for any space Y) coincides with the space $C(Y, X)$, endowed with the limitation topology introduced in Section 2.4.1.

Several common concepts in the metrizable case have to be modified for the nonmetrizable case. For instance, we call a closed subset A of a space X a *Z_τ-set* if the set

$$\{f \in C_\tau(X, X)\colon\ f(X) \cap A = \emptyset\}$$

is dense in the space $C_\tau(X, X)$. Similarly, elements of the closure (in $C_\tau(X, Y)$) of the set of homeomorphisms of X onto Y are called *τ-near-homeomorphisms*.

Interestingly enough, it is much easier to detect Z_τ-sets in the nonmetrizable case. For instance, every metrizable or finite-dimensional compactum in the Tychonov cube I^τ is a Z_τ-set [60].

5.2. *I^τ-manifolds*

In this section we assume that τ is an uncountable cardinal. A Lindelöff space X is an *I^τ-manifold* if it has a countable functionally open cover, each element of which is homeomorphic to a functionally open subspace of the Tychonov cube I^τ. A characterization of I^τ-manifolds was obtained by Ščepin [128]. First we need to extend Definition 2.5 to the nonmetrizable case.

DEFINITION 5.4. For a given complex L we say that a (locally) compact space X is *strongly τ-$[L]$-universal for compact spaces* if for any (locally) compact space B such that $\text{ed}(B) \leqslant [L]$ and $w(B) \leqslant \tau$, its closed subspace A, any (proper) map $f : B \to X$ such that the restriction f/A is a Z_τ-embedding, and any neighborhood U of f in $C_\tau(B, X)$, there is a Z_τ-embedding $g \in U$ such that $g/A = f/A$.

THEOREM 5.6 (Characterization of I^τ-manifolds, [51, 128]). *The following conditions are equivalent for a locally compact and Lindelöff* ANE-*space X of weight τ*:

(i) *X is an I^τ-manifold.*

(ii) *X is strongly τ-$[\{\text{pt}\}]$-universal for compact spaces.*

(iii) *For each compact space B of weight $< \tau$, the set of embeddings is dense in the space $C_\tau(B, X)$.*

(iv) *The set of embeddings is dense in the space $C_\tau(\{0, 1\}, X)$.*

(v) *X is homogeneous with respect to the pseudo-character, i.e., $\varphi(x, X) = \tau$ for each point $x \in X$.*

Condition (iii) in this theorem is the main difference between the metrizable and nonmetrizable cases. It implies statements which are not even true for $\tau = \omega$. For instance, if the product of two compact spaces is homeomorphic to I^τ, then one of them itself must be a copy of I^τ. Compare with Example 3.1. Another strange fact: the cone $\mathrm{Cone}(I^\tau)$ is not homeomorphic to I^τ (the vertex of the cone is a G_δ-point violating condition (v)), whereas $\mathrm{Cone}(I^\omega) \approx I^\omega$ (Theorem 3.6).

The following statement provides a natural bridge between the metrizable and non-metrizable cases.

THEOREM 5.7 ([128]). *The following conditions are equivalent for any space X:*
(i) *X is an I^τ-manifold.*
(ii) *There is an I^ω-manifold (unique up to a homeomorphism) Y such that $X \approx Y \times I^\omega$.*

This theorem has several corollaries.

THEOREM 5.8 (Stability of I^τ-manifolds, [128]). *If X is an I^τ-manifold, then $X \approx X \times I^\tau$.*

THEOREM 5.9 (Triangulation of I^τ-manifolds, [128]). *Every I^τ-manifold is triangulable, i.e., can be represented as the product $K \times I^\tau$ for some locally compact polyhedron K.*

THEOREM 5.10 (Open embedding theorem, [51]). *For any I^τ-manifold X, the product $X \times [0,1)$ is homeomorphic to a functionally open subspace of I^τ.*

THEOREM 5.11 (Homotopy classification, [128]). *I^τ-manifolds X and Y are homotopy equivalent if and only if the products $X \times [0,1)$ and $Y \times [0,1)$ are homeomorphic.*

THEOREM 5.12 (ANE-theorem, [128]). *The product $X \times I^\tau$ is an I^τ-manifold if and only if X is a locally compact and Lindelöff* ANE-*space.*

THEOREM 5.13 (Global Z_τ-set unknotting, [51, 60]). *Suppose that the homeomorphism $g: A \to B$ between Z_τ-sets of the compact I^τ-manifold X is homotopic to the inclusion map $A \hookrightarrow X$ ($\tau > \omega$). Then there is an autohomeomorphism $G: X \to X$ which extends g and is homotopic to* id_X.

THEOREM 5.14 (Local Z_τ-set unknotting, [51, 60]). *Let X be a compact I^τ-manifold, $\tau > \omega$. For each collection $\{\mathcal{U}_t\colon t \in T\} \subseteq \mathrm{cov}(X)$, $|T| < \tau$, there exists another collection $\{\mathcal{V}_t\colon t \in T'\}$, $|T'| < \tau$, such that the following condition is satisfied:*

(∗) *For any homeomorphism $h: A \to B$ between arbitrary Z_τ-sets of X which is $\{\mathcal{V}_t\colon t \in T'\}$-close to the inclusion map $A \hookrightarrow X$, there is a homeomorphism $H: X \to X$ such that $H/Z = h$ and which is $\{\mathcal{U}_t\colon t \in T\}$-close to* id_X.

THEOREM 5.15 ([51]). *Each compactum of weight τ, admitting a soft map onto the Tychonov cube I^τ, is homeomorphic to I^τ.*

It is not hard to see that the group $\mathbb{A}_p$ acts on the Hilbert cube I^ω, fixing a single point and acting freely off that point [88]. Hence $\mathbb{A}_p$ acts freely on the noncompact I^ω-manifold $I^\omega - \{\mathrm{pt}\} \approx I^\omega \times [0, 1)$. Are there *compact* I^ω-manifolds admitting a free action of the group $\mathbb{A}_p$? This question (R.D. Edwards [88]) is a weaker version of the well-known Hilbert–Smith conjecture[14], and to the best of my knowledge still remains open. The following statement (the proof of which is based on the spectral technique) answers a nonmetrizable version of Edwards' question.

THEOREM 5.16 ([51]). *There is no nonmetrizable* ANE-*compactum admitting a nontrivial semi-free action of any nonmetrizable compact group.*

5.3. R^τ-*manifolds*

A space X is an R^τ-*manifold* if it has a countable functionally open cover each element of which is homeomorphic to R^τ. A characterization of R^τ-manifolds was obtained in [41]. We start with the nonmetrizable counterpart of Definition 2.8.

DEFINITION 5.5. For a given complex $[L]$ we say that a realcompact space X is *strongly* τ-$[L]$-*universal for realcompact spaces* if for any realcompact space B such that $\mathrm{ed}(B) \leqslant [L]$ and $R\text{-}w(B) \leqslant \tau$, its closed subspace A, any map $f : B \to X$ such that the restriction f/A is a Z_τ-embedding, and any neighborhood U of f in $C_\tau(B, X)$, there is a Z_τ-embedding $g \in U$ such that $g/A = f/A$.

THEOREM 5.17 (Characterization of R^τ-manifolds, [51, 128]). *The following conditions are equivalent for any* ANE-*space X of weight τ:*

(i) *X is an R^τ-manifold.*
(ii) *X is strongly τ-$[\{\mathrm{pt}\}]$-universal for realcompact spaces.*
(iii) *For each realcompact space B of R-weight $< \tau$, the set of C-embeddings is dense in the space $C_\tau(B, X)$.*
(iv) *The set of embeddings is dense in the space $C_\tau(\omega, X)$.*

Condition (iii) in this theorem is the main difference between the metrizable and nonmetrizable cases. It implies statements which are not even true for $\tau = \omega$. For instance, if the product of two spaces is homeomorphic to R^τ, then one of them must be a copy of R^τ. Compare with Example 3.2.

The following statement provides a natural bridge between the metrizable and nonmetrizable cases.

THEOREM 5.18 ([51]). *The following conditions are equivalent for any space X:*

(i) *X is an R^τ-manifold.*
(ii) *There is an R^ω-manifold (unique up to a homeomorphism) Y such that $X \approx Y \times R^\omega$.*

[14] The Hilbert–Smith conjecture asks whether a locally compact group acting effectively on an n-manifold must be a Lie group.

This theorem has several corollaries.

THEOREM 5.19 (Stability of R^τ-manifolds, [51]). *If X is an R^τ-manifold, then $X \approx X \times R^\tau$.*

THEOREM 5.20 (Triangulation of R^τ-manifolds, [51]). *Every R^τ-manifold is triangulable, i.e., can be represented as the product $K \times R^\tau$ for some locally compact polyhedron K.*

THEOREM 5.21 (Open embedding theorem, [51]). *The following conditions are equivalent for any space X:*

(i) *X is an R^τ-manifold.*

(ii) *X is homeomorphic to a functionally open subspaces of R^τ.*

THEOREM 5.22 (Homotopy classification, [51]). *Homotopy equivalent R^τ-manifolds are homeomorphic.*

THEOREM 5.23 (ANE-theorem, [51]). *The product $X \times R^\tau$ is an R^τ-manifold if and only if X is an* ANE*-space.*

THEOREM 5.24 ([51]). *Every space of weight τ, admitting a soft map onto R^τ, is homeomorphic to R^τ.*

We conclude this section with the following generalization of Corollary 3.1.

THEOREM 5.25 ([47]). *Let τ be an uncountable cardinal. Then the following conditions are equivalent for any locally convex linear topological space X of weight τ:*

(i) *X is homeomorphic to R^τ.*

(ii) *X is an* AE*-space.*

(iii) *X is an* AE(0)*-space (recall that* $\mathrm{AE}(0) = \mathrm{AE}([S^0])$).

6. Topological groups

Quite often, simple topological assumptions in the presence of an additional algebraic structure have surprisingly strong consequences. We have seen examples of such situations in Corollary 3.1 and Theorem 5.25. Two more examples are provided in the following statements.

PROPOSITION 6.1 (Pontryagin–Haydon, [103, 124]). *Every compact group is an* AE(0)*-space.*

PROPOSITION 6.2 ([17]). *The following conditions are equivalent for a zero-dimensional topological group G:*

(a) *G is topologically equivalent to the product $(\mathbb{Z}_2)^\tau \times \mathbb{Z}^\kappa$.*

(b) *G is an* AE(0)*-space.*

Below we consider compact topological groups from the point of view of the general theory of absolute extensors.

6.1. *Abelian case – the Whitehead problem and characterization of Torus groups*

The problem of the topological characterization of torus groups (i.e., powers of the circle group $\mathbb{T}$) has been of considerable importance in the theory of compact abelian groups. The following statement collects most of the known information.

PROPOSITION 6.3. *Let G be a connected compact abelian group. Then the following conditions are equivalent*:

(1) *G is arcwise connected.*
(2) *G is locally arcwise connected.*
(3) *The map* $\exp_G : L(G) \to G$ *is surjective.*
(4) *The map* $\exp_G : L(G) \to G$ *is open.*
(5) $\mathrm{Ext}(\widehat{G}, \mathbb{Z}) = 0$, *i.e., the character group $\widehat{G}$ of G is a Whitehead group.*
(6) $\mathrm{Ext}(\widehat{G}, \mathbb{Z}) = 0$ *and each pure subgroup of $\widehat{G}$ of finite rank splits.*

In addition, if G is metrizable, *then the above conditions are equivalent to each of the following*:

(7) *$\widehat{G}$ is free.*
(8) *$\widehat{G}$ is $\aleph_1$-free, i.e., every countable subgroup of $\widehat{G}$ is free.*
(9) *$\widehat{G}$ is projective.*
(10) *G is injective (in the category of compact abelian groups).*
(11) *G is locally connected.*
(12) *G is a torus.*

In the nonmetrizable case situation is quite different. For instance, the equivalence (8) $\Leftrightarrow$ (11) remains true, whereas (11) $\Leftrightarrow$ (12) does not. The equivalence (5) $\Leftrightarrow$ (7) is undecidable in ZFC [130]. In topological terms this means that:

(i) There are (necessarily nonmetrizable) connected and locally connected compact abelian groups which are not tori.
(ii) It is undecidable within ZFC that the arcwise connectedness of a nonmetrizable compact abelian group forces the group to be isomorphic to a torus.

The following result provides a topological characterization of torus groups.

THEOREM 6.1 ([54]). *The following conditions are equivalent for a compact abelian group G*:

(i) *G is a torus group (both in the topological and algebraic senses).*
(ii) *G is an* AE(1)*-compactum.*
(iii) *The exponential map* $\exp_G : L(G) \to G$ *is* 0*-soft.*

Since, by duality, any statement in the category of compact abelian groups generates an equivalent statement in the category of abelian groups (see Proposition 6.3), it might be worth examining Theorem 6.1 from this point of view. By the equivalence (7) $\Leftrightarrow$ (12)

in Proposition 6.3, the statement corresponding to Theorem 6.1 in the category of abelian groups must provide a characterization of free groups. Since $\dim G = \operatorname{rank}(\widehat{G})$ for a finite dimensional compact abelian group G, and since the concept of AE(1) is defined in terms of 1-dimensional testing compacta, this would lead us to the concept of 1-projectivness (restrict, in the standard definition of projectivness, domains of testing epimorphisms by groups of rank 1) for abelian groups, and to the hope of a subsequent characterization of free abelian groups in these terms. The so obtained version of AE(1) will differ from the original topological concept of AE(1) and will not be a replacement for the latter in Theorem 6.1. The reason is simple – every 1-dimensional connected compact (abelian) group is metrizable and, consequently, the changed version of AE(1) would essentially coincide with arcwise connectedness. This shows that Theorem 6.1 provides a purely *topological* characterization of tori.

6.2. *Non-abelian case*

Extension theory allows us to characterize simple, connected and simply connected Lie groups. Here we present corresponding statements. Note that every connected Lie group topologically is an AE(1)-space and every connected and simply connected Lie group is an AE(2)-space.

THEOREM 6.2 ([52]). *The following conditions are equivalent for a compact group G:*
(i) *G is a simply connected* AE(1)*-compactum.*
(ii) *G is an* AE(2)*-compactum.*
(iii) *G is an* AE(3)*-compactum.*
(iv) *G is a product of simple, connected and simply connected compact Lie groups.*

This statement immediately implies

COROLLARY 6.1 ([52]). *The following conditions are equivalent for a nontrivial compact group G:*
(i) *G is an* AE(2)*-group with $\pi_3(G) = \mathbb{Z}$.*
(ii) *G is a simple, connected and simply connected Lie group.*

Note that the implication (ii) $\Rightarrow$ (i) in Corollary 6.1 is due to R. Bott (see [119, Theorem 3.8]).

References

[1] A.D. Aleksandrov, *Additive set-functions in abstract spaces*, Mat. Sbornik **8** (1940), 307–348; **9** (1941), 563–628; **13** (1943), 169–238.
[2] R.D. Anderson, *A characterization of the universal curve and a proof of its homogeneity*, Ann. Math. **67** (1958), 313–324.
[3] R.D. Anderson, *One-dimensional continuous curves and a homogeneity theorem*, Ann. Math. **68** (1958), 1–16.

[4] R.D. Anderson, *Hilbert space is homeomorphic to the countable infinite product of lines*, Bull. Amer. Math. Soc. **72** (1966), 515–519.
[5] R.D. Anderson, *On topological infinite deficiency*, Michigan Math. J. **126** (1967), 365–383.
[6] R.D. Anderson, *Topological properties of the Hilbert cube and the infinite product of open intervals*, Trans. Amer. Math. Soc. **126** (1967), 200–216.
[7] R.D. Anderson and T.A. Chapman, *Extending homeomorphisms to Hilbert cube manifolds*, Pacific J. Math. **38** (1971), 281–293.
[8] R.D. Anderson, D.W. Curtis and J. van Mill, *A fake topological Hilbert space*, Trans. Amer. Math. Soc. **272** (1982), 311–321.
[9] R.D. Anderson and J.D. McCharen, *On extending homeomorphisms to Frechet manifolds*, Proc. Amer. Math. Soc. **25** (1970), 283–289.
[10] R.D. Anderson and R.M. Schori, *Factors of infinite-dimensional manifolds*, Trans. Amer. Math. Soc. **142** (1969), 315–330.
[11] J. Baars, H. Gladdines and J. van Mill, *Absorbing systems in infinite-dimensional manifolds*, Topology Appl. **50** (1993), 147–182.
[12] T.O. Banakh, *Characterization of convex Z-sets in linear metric spaces*, Funct. Anal. Appl. **28** (1994), 285–286.
[13] T.O. Banakh and R. Cauty, *Topological classification of spaces of probability measures on co-analytic spaces*, Math. Notes **55** (1994), 8–13.
[14] T. Banakh, T. Radul and M. Zarichnyi, *Absorbing Sets in Infinite-Dimensional Manifolds*, VNTL Publishers, Lviv (1996).
[15] H.J. Baues, *Homotopy types*, Handbook of Algebraic Topology, I.M. James, ed., North-Holland, Amsterdam (1995), 1–72.
[16] H.J. Baues, *Combinatorial Homotopy and 4-Dimensional Complexes*, De Gruyter, Berlin (1991).
[17] M. Bell and A. Chigogidze, *Topological groups homeomorphic to products of discrete spaces*, Topology Appl. **53** (1993), 67–73.
[18] C. Bessaga and A. Pełczyński, *The estimated extension theorem, homogeneous collections and skeletons and their applications to the topological classification of linear metric spaces and convex sets*, Fund. Math. **69** (1970), 153–190.
[19] C. Bessaga and A. Pełczyński, *Selected Topics in Infinite-Dimensional Topology*, PWN, Warszawa (1975).
[20] M. Bestvina, *Characterizing k-dimensional universal Menger compacta*, Mem. Amer. Math. Soc. **71** (380) (1988).
[21] M. Bestvina, P.L. Bowers, J. Mogilski and J. Walsh, *Characterizing Hilbert space manifolds revisited*, Topology Appl. **24** (1986), 53–69.
[22] M. Bestvina and J. Mogilski, *Characterizing certain incomplete infinite-dimensional absolute retracts*, Michigan Math. J. **33** (1986), 291–313.
[23] K. Borsuk, *Über Isomorphie der Funktionalräume*, Bull. Int. Acad. Polon. Sci. et Lettres (1933), 1–10.
[24] K. Borsuk, *Theory of Retracts*, PWN, Warszawa (1967).
[25] K. Borsuk, *Theory of Shape*, PWN, Warszawa (1975).
[26] K. Borsuk and S. Ulam, *On symmetric products of topological spaces*, Bull. Amer. Math. Soc. **37** (1931), 875–882.
[27] H.G. Bothe, *Eine Einbettung m-dimensionalen Mengen in einen $(m+1)$-dimensionalen absoluten Retrakt*, Fund. Math. **52** (1963), 209–224.
[28] W. Browder, J. Levine and G.R. Livesay, *Finding a boundary for an open manifold*, Amer. J. Math. **87** (1965), 1017–1028.
[29] R. Cauty, *Une caractérisation des rétractes absolus de voisinage*, Fund. Math. **144** (1994), 11–22.
[30] R. Cauty, *Un espace métrique linéaire qui n'est las un rétracte absolu*, Fund. Math. **146** (1994), 85–99.
[31] R. Cauty, T. Dobrowolski, H. Gladdines and J. van Mill, *Les hyperspaces des retractes absolus et des retractes absolus de voisinage du plan*, Fund. Math. **148** (1995), 257–282.
[32] T.A. Chapman, *Dense sigma-compact subsets of infinite-dimensional manifolds*, Trans. Amer. Math. Soc. **154** (1971), 399–426.
[33] T.A. Chapman, *On some applications of infinite-dimensional manifolds to the theory of shape*, Fund. Math. **76** (1972), 181–193.
[34] T.A. Chapman, *On the structure of Hilbert cube manifolds*, Compositio Math. **24** (1972), 329–353.

[35] T.A. Chapman, *Cell-like mappings of Hilbert cube manifolds: Solution of a handle problem*, Topology Appl. **5** (1975), 123–145.
[36] T.A. Chapman, *Lectures on Hilbert Cube Manifolds*, Regional Series in Math., Vol. 28, Amer. Math. Soc., Providence, RI (1976).
[37] T.A. Chapman, *Simple homotopy theory of ANR's*, Topology Appl. **7** (1977), 165–174.
[38] T.A. Chapman and S. Ferry, *Obstructions to finiteness in the proper category*, Unpublished manuscript.
[39] T.A. Chapman and L.C. Siebenmann, *Finding a boundary for a Hilbert cube manifolds*, Acta Mathematica **137** (1976), 171–208.
[40] A. Chigogidze, *Trivial bundles with Tychonov cube fiber*, Mat. Notes **39** (1986), 747–756 (Russian).
[41] A. Chigogidze, *Noncompact absolute extensors in dimension n, n-soft mappings and their applications*, Math. USSR Izv. **28** (1987), 151–174.
[42] A. Chigogidze, *On the structure of non-metrizable AE(0)-spaces*, Mat. Notes **41** (1987), 406–411 (Russian).
[43] A. Chigogidze, *Compacta lying in the n-dimensional universal Menger compactum and having homeomorphic complements in it*, Math. USSR Sb. **61** (1988), 471–484.
[44] A. Chigogidze, *The theory of n-shapes*, Russian Math. Surveys **44** (1989), 145–174.
[45] A. Chigogidze, *Trivial bundles and near homeomorphisms*, Fund. Math. **132** (1989), 89–98.
[46] A. Chigogidze, *n-shapes and n-cohomotopy groups of compacta*, Math. USSR Sb. **66** (1990), 329–342.
[47] A. Chigogidze, *Locally convex linear topological spaces that are homeomorphic to the powers of the real line*, Proc. Amer. Math. Soc. **113** (1991), 599–609.
[48] A. Chigogidze, *Classification theorem for Menger manifolds*, Proc. Amer. Math. Soc. **116** (1992), 825–832.
[49] A. Chigogidze, *UV^n-equivalence and n-equivalence*, Topology Appl. **45** (1992), 283–291.
[50] A. Chigogidze, *Finding a boundary for a Menger manifold*, Proc. Amer. Math. Soc. **121** (1994), 631–640.
[51] A. Chigogidze, *Inverse Spectra*, North-Holland, Amsterdam (1996).
[52] A. Chigogidze, *Compact groups and absolute extensors*, Topology Proc. **22** (1997), 63–70
[53] A. Chigogidze, *Cohomological dimension of Tychonov spaces*, Topology Appl. **79** (1997), 197–228.
[54] A. Chigogidze, *Topological characterization of torus groups*, Topology Appl. **94** (1999), 13–25.
[55] A. Chigogidze, *Compactifications and universal spaces in extension theory*, Proc. Amer. Math. Soc. **128** (2000), 2187–2190.
[56] A. Chigogidze and V.V. Fedorchuk, *On some dimensional properties of 4-manifolds*, Topology Appl. **107** (2000), 67–78.
[57] A. Chigogidze, K. Kawamura and R.B. Sher, *Finiteness results in n-shape theory*, Topology Appl. **74** (1996), 3–16.
[58] A. Chigogidze, K. Kawamura and E.D. Tymchatyn, *Menger manifolds*, Continua with the Houston Problem Book, Lecture Notes in Pure and Appl. Math., Vol. 170 (1995), 37–88.
[59] A. Chigogidze, K. Kawamura and E.D. Tymchatyn, *Nöbeling spaces and pseudo-interiors of Menger compacta*, Topology Appl. **68** (1996), 33–65.
[60] A. Chigogidze and J.R. Martin, *Fixed point sets of Tychonov cubes*, Topology Appl. **64** (1995), 201–218.
[61] A. Chigogidze and M. Zarichnyi, *On absolute extensors modulo a complex*, Topology Appl. **86** (1998), 169–178.
[62] M.M. Cohen, *A Course in Simple-Homotopy Theory*, Springer, Berlin (1970).
[63] D.W. Curtis, *Hyperspaces homeomorphic to Hilbert space*, Proc. Amer. Math. Soc. **75** (1979), 126–130.
[64] D.W. Curtis and R.M. Schori, *Hyperspaces of Peano continua are Hilbert cubes*, Fund. Math. **101** (1978), 19–38.
[65] J.J. Dijkstra, J. van Mill and J. Mogilski, *Classification of finite-dimensional pseudo-boundaries and pseudo-interiors*, Trans. Amer. Math. Soc. **332**, (1992), 693–709.
[66] T. Dobrowolski, W. Marciszewski and J. Mogilski, *On topological classification of function spaces $C_p(X)$ of low Borel complexity*, Trans. Amer. Math. Soc. **328** (1991), 307–324.
[67] T. Dobrowolski and J. Mogilski, *Problems on topological classification of incomplete metric spaces*, Open Problems in Topology, J. van Mill and G.M. Reed, eds, North-Holland, New York (1990), 409–429.
[68] T. Dobrowolski and L.R. Rubin, *The hyperspace of infinite-dimensional compacta for covering and cohomological dimension are homeomorphic*, Pacific J. Math. **164** (1994), 15–39.
[69] T. Dobrowolski and L.R. Rubin, *The space of ANR's in R^n*, Fund. Math. **146** (1994), 31–58.

[70] T. Dobrowolski and K. Sakai, *Spaces of measures on metrizable spaces*, Topology Appl. **72** (1996), 215–258.
[71] T. Dobrowolski and H. Toruńczyk, *On metric linear spaces homeomorphic to l_2 and compact convex sets homeomorphic to Q*, Bull. Acad. Polon. Sci. **27** (1979), 883–887.
[72] A.N. Dranishnikov, *Absolute extensors in dimension n and n-soft mappings raising dimension*, Uspekhi Mat. Nauk **39** (1984), 55–95.
[73] A.N. Dranishnikov, *Universal Menger compacta and universal mappings*, Mat. Sbornik **129** (1986), 121–139.
[74] A.N. Dranishnikov, *On a problem of P. S. Alexandrov*, Mat. Sbornik **135** (1988), 551–557.
[75] A.N. Dranishnikov, *Extension of mappings into CW-complexes*, Mat. Sbornik **182** (1991), 47–56.
[76] A.N. Dranishnikov, *The Eilenberg–Borsuk theorem for maps in an arbitrary complex*, Mat. Sbornik **185** (1994), 81–90.
[77] A.N. Dranishnikov, *On extension theory of compact space*, Preprint (1998).
[78] A.N. Dranishnikov, J. Dydak and J. Keesling, *Extension of mappings*, Preprint.
[79] A.N. Dranishnikov and D. Repovš, *Cohomological dimension with respect to perfect groups*, Topology Appl. **74** (1996), 123–140.
[80] A.N. Dranishnikov and E.V. Ščepin, *Cell-like mappings*, Uspekhi Mat. Nauk **41** (1986), 49–90.
[81] J. Dugundji, *An extension of Tietze's theorem*, Pacific J. Math. **1** (1951), 353–367.
[82] J. Dydak, *Cohomological dimension and metrizable spaces. II*, Trans. Amer. Math. Soc. **348** (1996), 1647–1661.
[83] J. Dydak, *Cohomological dimension theory*, Handbook of Geometric Topology, R.J. Davermann and R.B. Sher, eds, Elsevier, Amsterdam (2002), 423–470.
[84] J. Dydak and J. Segal, *Shape Theory*, Lecture Notes Math., Vol. 688 (1978).
[85] J. Dydak and J. Segal, *A list of open problems in shape theory*, Open Problems in Topology, J. van Mill and G.M. Reed, eds, North-Holland, New York (1990), 457–467.
[86] J. Dydak and J. Walsh, *Spaces without cohomological dimension preserving compactifications*, Proc. Amer. Math. Soc. **113** (1991), 1155–1162.
[87] J. Dydak and J. Walsh, *Infinite-dimensional compacta having cohomological dimension two: An application of the Sullivan conjecture*, Topology **32** (1993), 93–104.
[88] R.D. Edwards, *Problems:* 1987 *Western Workshop in Geometric Topology*, Oregon State Univ. Press (1987), 46–48.
[89] R.D. Edwards, *Characterizing infinite-dimensional manifolds topologically (after Henryk Toruńczyk)*, Lect. Notes Math., Vol. 540 (1999), 278–302.
[90] D.A. Edwards and R. Geoghegan, *Shapes of complexes, ends of manifolds, homotopy limits and the Wall obstruction*, Ann. of Math. **101** (1975), 521–535.
[91] V.V. Fedorchuk, *Covariant functors in the category of compacta, absolute retracts and Q-manifolds*, Russian Math. Surveys **36** (3) (1981), 211–233.
[92] V.V. Fedorchuk, *On open mappings*, Uspekhi Mat. Nauk **37** (1982), 187–188 (Russian).
[93] S. Ferry, *The homeomorphism group of a compact Hilbert cube manifold is an ANR*, Ann. Math. **106** (1977), 101–119.
[94] S. Ferry, *UV^k-equivalent compacta*, Lecture Notes Math., Vol. 1283 (1987), 88–114.
[95] R.H. Fox, *On the Lusternik–Schnirelman category*, Ann. of Math. **42** (1941), 333–370.
[96] R. Geoghegan, *Open problems in infinite-dimensional topology*, Topology Proc. **4** (1979), 287–330.
[97] R. Geoghegan and W. Haver, *On the space of piecewise linear homeomorphisms of a manifold*, Proc. Amer. Math. Soc. **55** (1976), 145–151.
[98] R. Geoghegan and R. Summerhill, *Pseudo-boundaries and pseudo-interiors in Euclidean space and topological manifolds*, Trans. Amer. Math. Soc. **194** (1974), 141–165.
[99] L. Gillman and M. Jerison, *Rings of Continuous Functions*, Van Nostrand, New York (1960).
[100] O. Hanner, *Some theorems on absolute neighborhood retracts*, Arkiv Mat. **1** (1951), 389–408.
[101] O. Hanner, *Retractions and extensions of mappings of metric and non-metric spaces*, Arkiv Mat. **2** (1952), 315–360.
[102] W.E. Haver, *Mappings between ANR's that are fine homotopy equivalences*, Pacific J. Math. **58** (1975), 457–461.

[103] R. Haydon, *On a problem of Pełczyński: Milutin spaces, Dugundji spaces and $AE(0 - \dim)$*, Studia Math. **52** (1974), 23–31.

[104] D.W. Henderson, *Open subsets of Hilbert space*, Compositio Math. **21** (1969), 312–318.

[105] D.W. Henderson, *Infinite-dimensional manifolds are open subsets of Hilbert space*, Topology **9** (1970), 25–34.

[106] D.W. Henderson and R.M. Schori, *Topological classification of infinite-dimensional manifolds by homotopy type*, Bull. Amer. Math. Soc. **76** (1970), 121–124.

[107] M.I. Kadec, *On topological equivalence of separable Banach spaces*, Soviet Math. Dokl. **7** (1966), 319–322.

[108] J. Keesling and D. Wilson, *The group of PL-homeomorphisms of a compact PL-manifold is an l_2^f-manifold*, Trans. Amer. Math. Soc. **193** (1975), 249–256.

[109] O.H. Keller, *Die Homeömorphie der compakten convexen Mengen im Hilbertschen Raum*, Math. Ann. **105** (1931), 748–758.

[110] V.L. Klee, *Some topological properties of convex sets*, Trans. Amer. Math. Soc. **78** (1955), 30–45.

[111] V.I. Kuzminov, *Homological dimension theory*, Russian Math. Surveys **23** (1968), 1–45.

[112] S. Lefshetz, *On compact spaces*, Ann. Math. **32** (1931), 521–538.

[113] R. Luke and W. Mason, *The space of autohomeomorphisms on a compact two-manifold is an absolute neighborhood retract*, Trans. Amer. Math. Soc. **164** (1972), 275–285.

[114] S. MacLane and J.H.C. Whitehead, *On the 3-type of a complex*, Proc. Nat. Acad. Sci. USA **36** (1950), 41–48.

[115] S. Mardešić and J. Segal, *Shape Theory*, North-Holland, Amsterdam (1982).

[116] K. Menger, *Allgemeine Raume und Cartesische Raume. Zweite Mitteilung: "Über umfas–sendste n-dimensional Mengen"*, Proc. Akad. Amsterdam **29** (1926), 1125–1128.

[117] J. van Mill, *A rigid space X for which $X \times X$ is homogeneous; an application of infinite-dimensional topology*, Proc. Amer. Math. Soc. **83** (1981), 597–600.

[118] J.W. Milnor, *On spaces having the homotopy type of a CW-complex*, Trans. Amer. Math. Soc. **90** (1959), 272–280.

[119] M. Mimura, *Homotopy theory of Lie groups*, Handbook of Algebraic Topology, North-Holland, Amsterdam (1995), 951–991.

[120] J. Mogilski, *Characterizing the topology of infinite dimensional σ-compact manifolds*, Proc. Amer. Math. Soc. **92** (1984), 111–118.

[121] G.M. Nepomnyashchii, *On the spectral decomposition of the multi-valued absolute retracts*, Uspekhi Mat. Nauk **36** (1981), 221–222 (Russian).

[122] W. Olszewski, *Completion theorem for cohomological dimensions*, Proc. Amer. Math. Soc. **123** (1995), 2261–2264.

[123] W. Olszewski, *Universal separable metrizable spaces of given cohomological dimension*, Topology Appl. **61** (1995), 293–299.

[124] L.S. Pontryagin, *Topological Groups*, Princeton Univ. Press, Princeton, NJ (1946).

[125] T. Porter, *Proper homotopy theory*, Handbook of Algebraic Topology, I.M. James, ed., North-Holland, Amsterdam (1995), 127–167.

[126] E.V. Ščepin, *Topology of limit spaces of uncountable inverse spectra*, Uspekhi Mat. Nauk **31** (1976), 191–226 (Russian).

[127] E.V. Ščepin, *Finite-dimensional bicompact absolute neighborhood retracts are metrizable*, Dokl. Akad. Nauk SSSR **233** (1977), 304–307.

[128] E.V. Ščepin, *On Tychonov manifolds*, Dokl. Akad. Nauk SSSR **246** (1979), 551–554 (Russian).

[129] R.M. Schori, *Topological stability for infinite-dimensional manifolds*, Compositio Math. **23** (1971), 87–100.

[130] S. Shelah, *Infinite Abelian groups, Whitehead problem and some consequences*, Israel J. Math. **18** (1974), 243–256.

[131] L.C. Siebenmann, *The obstruction to finding a boundary for an open manifold of dimension greater than five*, Doctoral dissertation, Princeton Univ. Press, Princeton, NJ (1965).

[132] L.C. Siebenmann, *L'invariance topologique du type simple d'homotopic*, Seminaire Bourbaki, 25^e anee, n° 428 (1972/73).

[133] W. Sierpinski, *Sur une courbe cantorienne qui contient une image biunivoque et continue de toute courbe donnee*, C. R. Acad. Paris **162** (1916), 629–632.
[134] E.H. Spanier, *Algebraic Topology*, McGraw-Hill, New York (1966).
[135] M.A. Štanko, *Solution of Menger's problem in the class of compacta*, Soviet Math. Dokl. **12** (1971), 1846–1849.
[136] H. Toruńczyk, Ph.D. Thesis, Warsaw.
[137] H. Toruńczyk, *Absolute retracts as factors of normed linear spaces*, Fund. Math. **86** (1974), 53–67.
[138] H. Toruńczyk, *Homeomorphism groups of compact Hilbert cube manifolds that are manifolds*, Bull. Pol. Acad. Sci. **25** (1977), 401–408.
[139] H. Toruńczyk, *Concerning locally homotopy negligible sets and characterization of l_2-manifolds*, Fund. Math. **101** (1978), 93–110.
[140] H. Toruńczyk, *On CE-images of the Hilbert cube and characterization of Q-manifolds*, Fund. Math. **106** (1980), 31–40.
[141] H. Toruńczyk, *Characterizing Hilbert space topology*, Fund. Math. **111** (1981), 247–262.
[142] H. Toruńczyk and J.E. West, *Fibrations and bundles with Hilbert cube manifold fibers*, Mem. Amer. Math. Soc. **80** (406) (1989).
[143] V.S. Varadarajan, *Measures on topological spaces*, Mat. Sbornik **55** (1961), 35–100.
[144] C.T.C. Wall, *Finiteness conditions for CW-complexes*, Ann. of Math. **81** (1965), 56–69.
[145] J.J. Walsh, *Dimension, cohomological dimension and cell-like mappings*, Lecture Notes Math., Vol. 870 (1981), 105–118.
[146] J.E. West, *Infinite products which are Hilbert cubes*, Trans. Amer. Math. Soc. **150** (1970), 1–25.
[147] J.E. West, *The ambient homeomorphy of an incomplete subspace of infinite-dimensional Hilbert spaces*, Pacific J. Math. **34** (1970), 257–267.
[148] J.E. West, *Mapping Hilbert cube manifolds to ANR's. A solution of a conjecture of Borsuk*, Ann. Math. **106** (1977), 1–18.
[149] J.E. West, *Open problems in infinite dimensional topology*, Open Problems in Topology, J. van Mill and G.M. Reed, eds, North-Holland, New York (1990), 523–597.
[150] J.H.C. Whitehead, *On the homotopy type of ANR's*, Bull. Amer. Math. Soc. **54** (1948), 1133–1145.
[151] J.H.C. Whitehead, *Combinatorial homotopy, I*, Bull. Amer. Math. Soc. **55** (1949), 213–245.
[152] G.T. Whyburn, *Topological characterization of Sierpinski curve*, Fund. Math. **45** (1958), 320–324.
[153] M. Wojdyslawski, *Retractes absolus et hyperspaces des continus*, Fund. Math. **32** (1939), 184–192.

CHAPTER 8

Nonpositive Curvature and Reflection Groups

Michael W. Davis*

Department of Mathematics, The Ohio State University, 231 W. 18th Avenue, Columbus, OH 43210, USA
E-mail: mdavis@math.ohio-state.edu

Contents

*Partially supported by NSF Grants DMS-9208071 and DMS-9505003.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Introduction

A space is *aspherical* if its universal cover is contractible. Examples of aspherical spaces occur in differential geometry (as complete Riemannian manifolds of nonpositive sectional curvature), in Lie groups (as $\Gamma\backslash G/K$ where G is a Lie group, K is a maximal compact subgroup and Γ is a discrete torsion-free subgroup), in 3-manifold theory and as certain 2-dimensional cell complexes. The main purpose of this paper is to describe another interesting class of examples coming from the theory of reflection groups (also called "Coxeter groups"). One of the main results is explained in Sections 1.3 and 2.3: given a finite simplicial complex L, there is a compact aspherical polyhedron X such that the link of each vertex in X is isomorphic to the barycentric subdivision of L. A version of this result first appeared in [29]. Later, in [45], Gromov showed that the polyhedron X can be given a piecewise Euclidean metric which is nonpositively curved in the sense of Aleksandrov. This gives a new proof of the asphericity of X (cf. Theorem 1.5 below). As we vary the choice of the link L we get polyhedra X with a variety of interesting properties. For example, if L is homeomorphic to an $(n-1)$-sphere, then X is an n-manifold. More examples are discussed in Section 2.5.

Section 1 covers background material on nonpositive curvature. The main examples are discussed in Section 2. Section 3 deals with some related aspherical complexes which arise in the study of complements of arrangements of hyperplanes.

This paper began as a set of notes for three lectures which I gave at the Eleventh Annual Workshop in Geometric Topology at Park City, Utah in June 1994. In the course of preparing it for this volume I have added approximately 25 percent more material, notably, Sections 1.4, 1.5, 2.4 and parts of Section 2.5. I would like to thank Lonette Stoddard for preparing the figures.

1. Nonpositively curved spaces

The notion of "nonpositive curvature" (or more generally of "curvature bounded above by a real number ε") makes sense for a more general class of metric spaces than Riemannian manifolds: one need only assume that any two points can be connected by a geodesic segment. For such spaces, the concept of curvature bounded above by ε can be defined via comparison triangles and the so-called "$CAT(\varepsilon)$-inequality". (This terminology is due to Gromov.) This is explained in Section1.1. In Section 1.2 we consider "piecewise constant curvature polyhedra" and give a condition (in terms of links of vertices) for such a polyhedron to have curvature bounded from above. The condition is that each link be $CAT(1)$. In Section 1.3 we discuss criteria for such a link to be $CAT(1)$. The two conditions we are most interested in are given in Gromov's Lemma and Moussong's Lemma. These give criteria for piecewise spherical simplicial complexes (with sufficiently big simplices) to be $CAT(1)$. In Section 1.4 we show that a proper $CAT(0)$ space can be compactified by adding an ideal boundary consisting of "endpoints" of geodesic rays. In this section we also discuss the notion of an "infinitesimal shadow" which measures the nonuniqueness of geodesic continuation at a point. In Section 1.5 we sketch a proof of Theorem 1.47 which asserts that the compactification of a $CAT(0)$ PL manifold is homeomorphic to a disk. It is also indicated how such a result might fail in the non-PL context. (Explicit examples of this failure are given in Section 2.5(g).) In Section 1.6 we discuss a conjecture of H. Hopf concerning the Euler characteristic of a closed, nonpositively curved, even-dimensional manifold. Using the combinatorial version of the Gauss–Bonnet Theorem this leads us to a conjecture concerning a number associated to a piecewise spherical structure on an odd-dimensional sphere.

1.1. *The $CAT(\varepsilon)$-inequality*

Given a smooth Riemannian manifold M one defines its "curvature tensor" and from this its "sectional curvature". The sectional curvature K of M is a real-valued function on the set of all pairs (x, P) where x is a point in M and P is a tangent 2-plane at x. Given a real number ε, we say that "M has curvature $\leqslant \varepsilon$" and write $K(M) \leqslant \varepsilon$ if the sectional curvature K is bounded above by ε.

It has long been recognized that the condition that the curvature of M is bounded above is equivalent to a condition which can be phrased purely in terms of the underlying metric (i.e., in terms of the distance function) on M. In fact, there are several possible versions of such a condition. We shall focus on one called the "$CAT(\varepsilon)$ condition" by Gromov. ("C" stands for either "Cartan" or "comparison", "A" of "Aleksandrov", and "T" for "Toponogov".) Once one has such a condition one can define the notion of "curvature $\leqslant \varepsilon$" for many "singular" metric spaces, that is, for a more general class of metric spaces than Riemannian manifolds.

Good references for this material are [44] and [15].

We begin by stating the following Comparison Theorem of Aleksandrov. A proof can be found in the article of M. Troyanov in [44].

THEOREM 1.1 (Aleksandrov). *Let M be a simply connected, complete Riemannian manifold and ε a real number. Then $K(M) \leqslant \varepsilon$ if and only if each geodesic triangle in M (of perimeter $\leqslant 2\pi/\sqrt{\varepsilon}$) satisfies the $CAT(\varepsilon)$-inequality.*

A "geodesic triangle" in M means a configuration in M consisting of three points (the "vertices") and three (minimal) geodesic segments connecting them (the "edges"). The term "$CAT(\varepsilon)$" is explained below.

As $\varepsilon > 0$, $= 0$, or < 0, let M^2_ε stand for S^2_ε (the 2-sphere of constant curvature ε) $\mathbb{E}^2$ (the Euclidean plane) or $\mathbb{H}^2_\varepsilon$ (the hyperbolic plane of curvature ε).

Let T be a geodesic triangle in M. A *comparison triangle* for T is a geodesic triangle T^* in M^2_ε with the same edge lengths as T. Choose a vertex x of T and a point y on the opposite edge. Let x^* and y^* denote the corresponding points in T^*. (See Figure 1.)

The $CAT(\varepsilon)$-inequality is

$$d(x, y) \leqslant d^*\big(x^*, y^*\big)$$

where d and d^* denote distance in M and M^2_ε, respectively.

REMARK. S^2_ε is the sphere of radius $1/\sqrt{\varepsilon}$. Since any geodesic triangle $T^* \subset S^2_\varepsilon$, must lie in some hemisphere, we see that the perimeter of T^* can be no larger than $2\pi/\sqrt{\varepsilon}$ (the

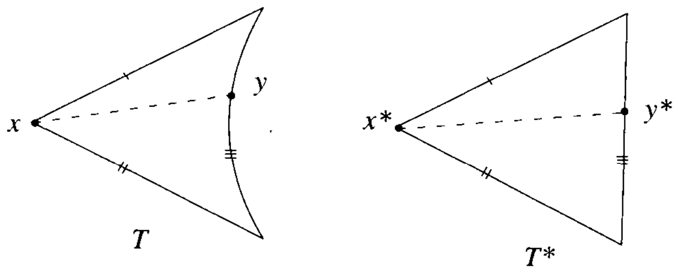


Fig. 1.

circumference of the equator). So, when $\varepsilon > 0$, the $CAT(\varepsilon)$-inequality only makes sense for triangles of perimeter $\leqslant 2\pi/\sqrt{\varepsilon}$.

Now let (X, d) be a metric space. A path $\alpha : [a, b] \to X$ is a *geodesic* if it is an isometric embedding, i.e., if $d(\alpha(t), \alpha(s)) = |t - s|$ for all s, t in $[a, b]$.

DEFINITION 1.2. A metric space X is a *geodesic space* (or a "length space") if an two points can be connected by a geodesic segment.

We shall also assume that X is complete and locally compact. (The hypothesis of local compactness could be removed, cf. [15].)

The notion of a geodesic triangle clearly makes sense in a geodesic space as does the $CAT(\varepsilon)$-inequality.

DEFINITION 1.3. A geodesic space X is $CAT(\varepsilon)$ if the $CAT(\varepsilon)$-inequality holds for all geodesic triangles T of perimeter $\leqslant 2\pi/\sqrt{\varepsilon}$ and for all choices of vertex x and point y on the opposite edge. (The condition that the perimeter be $\leqslant 2\pi/\sqrt{\varepsilon}$ is interpreted to be vacuous if $\varepsilon \leqslant 0$.) X has *curvature* $\leqslant \varepsilon$, written $K(X) \leqslant \varepsilon$, if it satisfies $CAT(\varepsilon)$ locally.

REMARKS. (i) If $\varepsilon' < \varepsilon$, then $CAT(\varepsilon')$ implies $CAT(\varepsilon)$ and $K(X) \leqslant \varepsilon'$ implies $K(X) \leqslant \varepsilon$.

(ii) There is a completely analogous definition of curvature bounded from below: one simply reverses the $CAT(\varepsilon)$-inequality. (See [1] or [52], Chapter 16 in this Handbook.)

Some consequences of $CAT(\varepsilon)$. (i) *There are no digons in X of perimeter $< 2\pi/\sqrt{\varepsilon}$.* (A *digon* is a configuration consisting of two distinct geodesic segments between points x and y.) The reason is that we could introduce a third vertex in the interior of one segment and obtain a triangle for which the $CAT(\varepsilon)$-inequality clearly fails. As special cases of this principle, we have the following.

(a) If X is $CAT(1)$, then a geodesic between two points of distance $< \pi$ is unique.

(b) If X is $CAT(1)$, then there is no closed geodesic of length $< 2\pi$. (A *closed geodesic* is an isometric embedding of a circle.)

(c) If X is $CAT(0)$, then any two points are connected by a unique geodesic.

(ii) *If X is $CAT(0)$, then the distance function $d : X \times X \to [0, \infty)$ is convex.* (In general, a function $\varphi : Y \to \mathbb{R}$ on a geodesic space Y is *convex*, if given any geodesic path

$\alpha : [a, b] \to Y$ the function $\varphi \circ \alpha : [a, b] \to \mathbb{R}$ is a convex function. In particular, $X \times X$, with the product metric, is a geodesic space and the statement that $d : X \times X \to [0, \infty)$ is convex means that given geodesic paths $\alpha : [a, b] \to X$ and $\beta : [c, d] \to X$ the function $(s, t) \to d(\alpha(s), \beta(t))$ is a convex function on $[a, b] \times [c, d]$.)

There is the following generalization of the Cartan–Hadamard Theorem for nonpositively curved manifolds.

PROPOSITION 1.4. *If X is a geodesic space with convex distance function (e.g., if X is $CAT(0)$), then X is contractible.*

PROOF. The convexity of the distance function implies that X has no digons. Hence, any two points of X are connected by a unique geodesic. Choose a base point x_0 and define the contraction $H : X \times I \to X$ by contracting along the geodesic to x_0. The proof that H is continuous follows easily from the convexity of d. □

REMARK. Suppose $K(X) \leqslant \varepsilon$. Then since $CAT(\varepsilon)$ holds locally, X is locally convex (i.e., in any sufficiently small open set, any two points are connected by a unique geodesic). Therefore, X is locally contractible. In particular, any such X has a universal cover.

THEOREM 1.5. *Let $\varepsilon \leqslant 0$. If X is a geodesic space with $K(X) \leqslant \varepsilon$, then its universal cover $\widetilde{X}$ is $CAT(\varepsilon)$. (In particular, $\widetilde{X}$ is contractible.)*

This theorem is stated by Gromov in [45, p. 119] and proved in W. Ballman's article in [44, p. 193]. (Quite possibly, it was known to Aleksandrov.)

REMARK. Theorem 1.5 is not true for $\varepsilon > 0$. There is an analogous result for $\varepsilon > 0$: the hypothesis of simple connectivity is unimportant, but one needs to rule out the possibility of closed geodesics of length $< 2\pi/\sqrt{\varepsilon}$. A version of this is stated as Lemma 1.13, below.

COROLLARY 1.6. *If $K(X) \leqslant 0$, then X is a $K(\pi, 1)$-space (i.e., X is aspherical).*

1.2. *Piecewise constant curvature polyhedra*

Let M^n_ε stand for S^n_ε, $\mathbb{E}^n$ or $\mathbb{H}^n_\varepsilon$ as ε is greater than, equal to, or less than 0, respectively. A "half-space" in S^n_ε is a hemisphere; a "half-space" in $\mathbb{E}^n$ or $\mathbb{H}^n_\varepsilon$ has its usual meaning.

DEFINITION 1.7. A (convex) *cell* in M^n_ε is a compact intersection of a finite number of half-spaces. (When $\varepsilon > 0$, one can also require that the cell does not contain a pair of antipodal points.)

DEFINITION 1.8. An *M_ε cell complex* X is a cell complex formed by gluing together cells in M^n_ε via isometries of their faces. (ε is fixed, n can vary.) If $\varepsilon = 0$, X is called *piecewise Euclidean* (abbreviated PE). If $\varepsilon = 1$, X is *piecewise spherical* (abbreviated PS). If $\varepsilon = -1$, X is *piecewise hyperbolic* (abbreviated PH).

EXAMPLE 1.9. The surface of a cube is a PE complex.

If X is an M_ε cell complex, then we can measure the length ℓ of a path in X: the length of the portion of the path within a given cell is defined using arc length in M_ε^n. The *intrinsic metric* ℓ on X is defined as follows:

$$d(x, y) = \inf\{\ell(w) \mid \alpha \text{ is a path from } x \text{ to } y\}.$$

(If X is not path connected, the d may take ∞ as a value.)

Does the intrinsic metric give X the structure of a geodesic space? The issue is whether the infimum occurring in the definition of d can actually be realized by a minimal path. If X is locally finite and if there is a $\delta > 0$ so that all closed δ-balls in X are compact (e.g., if X is a finite complex), then the Arzela–Ascoli Theorem implies that X is a complete geodesic space.

Links. Suppose that σ is an n-cell in M_ε^n and that v is a vertex of σ. The Riemannian metric on M_ε^n gives an inner product on its tangent space $T_v(M_\varepsilon^n)$ at v. The set of inward pointing directions at v is subset of the unit sphere in $T_v(M_\varepsilon^n)$. In fact, this subset is a spherical cell, which we denote by $Lk(v, \sigma)$. We think of it as a cell in S^{n-1}, well-defined up to isometry. (See Figure 2.)

If X is an M_ε-complex and v is a vertex of X, then the *link of v in X* is defined by

$$Lk(v, X) = \bigcup_{\substack{v \\ v \subset \sigma}} Lk(v, \sigma).$$

This is a PS cell complex. Thus, the link of a vertex in any M_ε cell complex has a natural piecewise spherical structure.

EXAMPLE 1.10. As in Example 1.9, let X be the surface of a cube and v, a vertex. Then the link of v in each square is a circular arc of length $\pi/2$; hence, the link of v in X is a circle of length $3\pi/2$.

In [45, p. 120] Gromov gave the following "infinitesimal" condition for deciding if a piecewise constant curvature polyhedron has curvature bounded from above.

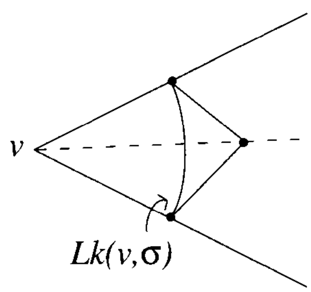


Fig. 2.

THEOREM 1.11 (Aleksandrov, Gromov, Ballman). *Suppose X is an M_ε-cell complex. Then $K(X) \leqslant \varepsilon$ if and only if for each vertex v, $Lk(v, X)$ is $CAT(1)$.*

A proof of this can be found in Ballman's article in [44, p. 197]. The result must have also been known to Aleksandrov's school, since they knew that an "M_ε cone" on a $CAT(1)$ space was $CAT(\varepsilon)$.

EXAMPLE 1.12. A PS structure on a circle S is $CAT(1)$ if and only if $\ell(S) \geqslant 2\pi$. Therefore, a PE structure on a surface has $K \leqslant 0$ if and only if at each vertex the sum of the angles is $\geqslant 2\pi$. For example, the surface of a cube does not have nonpositive curvature.

1.3. *The $CAT(1)$ condition for links*

In order to use Theorem 1.11 we need to be able to decide when the link of a vertex is $CAT(1)$. So, suppose L is some PS cell complex. We need to be able to answer the following.

QUESTION. How do you tell if L is $CAT(1)$?

The following lemma gives an inductive procedure for studying this question.

LEMMA 1.13. *A PS complex L is $CAT(1)$ if and only if*
(i) *$K(L) \leqslant 1$, and*
(ii) *every closed geodesic in L has length $\geqslant 2\pi$.*

By Theorem 1.11, condition (i) can be checked by looking at links of vertices in L. Thus, (ii) is the crucial condition.

We next would like to explain several situations in which we have a satisfactory answer to our question. These will be grouped under the following headings:
(a) Gromov's Lemma,
(b) Moussong's Lemma,
(c) Orthogonal joins,
(d) Spherical buildings,
(e) Polar duals of hyperbolic cells,
(f) Branched covers of round spheres.
In these notes we will mostly be concerned with the first two headings (and we will confine ourselves to a few brief comments about the other four).

(a) *Gromov's Lemma.* Let $\square^n$ denote a regular n-cube in $\mathbb{E}^n$ and let v be a vertex of $\square^n$. Then $Lk(v, \square^n)$ is the regular spherical $(n-1)$-simplex Δ^{n-1} spanned by the standard basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$. (So, Δ^{n-1} is the intersection of S^{n-1} with the positive "quadrant" $[0, \infty)^n$ in $\mathbb{R}^n$.)

A spherical $(n-1)$-simplex isometric to Δ^{n-1} will be called an *all right* simplex.

An all right simplex is characterized by the fact that all its edge lengths are $\pi/2$. Alternatively, it can be characterized by the fact that all its dihedral angles are $\pi/2$.

DEFINITION 1.14. A PS simplicial cell complex is *all right* if each of its simplices is all right.

EXAMPLE 1.15. If X is a PE cubical complex, then each of its links is an all right simplicial cell complex.

DEFINITION 1.16. A simplicial complex K is a *flag complex* if any finite set of vertices, which are pairwise joined by edges, span a simplex in K.

Combinatorialists use "clique complex" instead of "flag complex". Alternative terminology, which has been used elsewhere, is that K is "determined by its 1-skeleton", or K has "no empty simplices", or K satisfies the "no Δ-condition". (The last is Gromov's terminology.) The term "flag complex" is taken from [17].

REMARK 1.17. Let V be a set with a symmetric and reflexive relation (an "incidence relation"). Let K be the abstract simplicial complex whose simplices are the finite subsets of V which are pairwise related. Then K is a flag complex. Conversely, given a flag complex K, one defines a relation on its vertex set V by saying that two vertices are related if they are joined by an edge. This relation gives back K as its associated complex.

EXAMPLE 1.18. Let $\mathcal{P}$ be a poset. Then if we make the order relation symmetric and take the associated simplicial complex, we get a flag complex. Its poset of simplices is denoted by $\mathcal{P}'$, and called the *derived complex* of $\mathcal{P}$. The elements of $\mathcal{P}'$ are finite chains $(v_0 < \cdots < v_k)$ in $\mathcal{P}$.

EXAMPLE 1.19. If $\mathcal{P}$ is the poset of cells in a cell complex, then $\mathcal{P}'$ can be identified with the poset of simplices in its barycentric subdivision. Thus, *the barycentric subdivision of any cell complex is a flag complex.*

EXAMPLE 1.20. If K is the boundary of an m-gon (i.e., K is a circle with m edges) then K is a flag complex if and only if $m > 3$.

LEMMA 1.21 (Gromov's Lemma). *Let L be an all right, PS simplicial complex. Then L is $CAT(1)$ if and only if it is a flag complex.*

COROLLARY 1.22 (Berestovskii [8]). *Any polyhedron has a PS structure which is $CAT(1)$.*

PROOF. Let L be a cell complex. By taking the barycentric subdivision we may assume that L is a flag complex. Then give L a piecewise spherical structure by declaring each simplex to be all right. □

COROLLARY 1.23. *Let X be a PE cubical complex. Then $K(X) \leqslant 0$ if and only if the link of each vertex is a flag complex.*

APPLICATION 1.24 (*Hyperbolization*). In [45] Gromov described several functorial procedures for converting a cell complex J (usually a simplicial or cubical complex) into a PE cubical complex $\mathcal{H}(J)$ with nonpositive curvature. (See also [35] and [21].) $\mathcal{H}(J)$ is called a "hyperbolization" of J. Since $\mathcal{H}(J)$ is aspherical it cannot, in general, be homeomorphic to J. However, there is a natural surjection $\mathcal{H}(J) \to J$. Also, $\mathcal{H}(J)$ should have the same local structure as J in the following sense: the link of each "hyperbolized cell" is PL homeomorphic to the link of the corresponding cell in J. Usually, the new link will be the barycentric subdivision of the old one (or else the suspension of a barycentric subdivision of an old link). Thus, the new links will be flag complexes and Gromov's Lemma can be used to prove that $\mathcal{H}(J)$ is nonpositively curved. (A different argument is given in [35] and [45].)

The proof of Gromov's Lemma is based on the following.

SUBLEMMA 1.25 ([45, p. 122]). *Let v be a vertex in an all right, PS simplicial complex and let B be the closed ball of radius $\pi/2$ about v (i.e., B is the closed star of v). Let x, y be points in ∂B (the sphere of radius $\pi/2$ about v) and let γ be a geodesic segment from x to y such that γ intersects the interior of B. Then $\ell(\gamma) \geqslant \pi$.*

PROOF. Let Δ be an all right simplex in B with one vertex at v and suppose that γ intersects the interior of Δ. Consider the union of all geodesic segment which start at v, pass through a point in $\gamma \cap \Delta$ and end on the face of Δ opposite to v. It is an isoceles spherical 2-simplex with two edges of length $\pi/2$ as in Figure 3. (Think of a spherical 2-simplex with one vertex at the north pole and the other two on the equator.) Let Ω be the union of all these 2-simplices. Then Ω can be "developed" onto the northern hemisphere of S^2. If $\ell(\gamma) < \pi$, then Ω is isometric to a region of S^2 so that v maps to the north pole and x and y to points on the equator. But if two points on the equator of S^2 are of distance $< \pi$, then the geodesic between them is a segment of the equator. This contradicts the hypothesis that the image of γ intersects the open northern hemisphere. □

PROOF OF GROMOV'S LEMMA. Let L be an all right, PS simplicial cell complex. First suppose that L is not a flag complex. Then either L or the link of some simplex of L

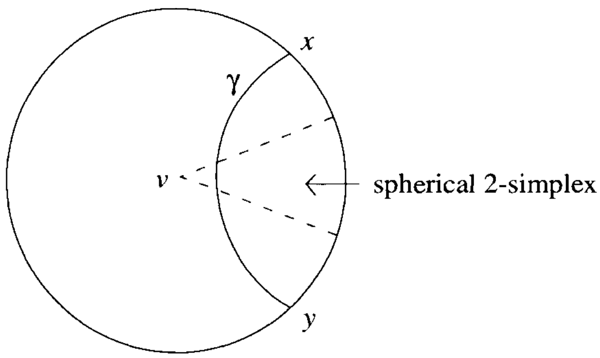


Fig. 3.

contains an "empty" triangle. Such a triangle is a closed geodesic of length $3\pi/2$ (which is $< 2\pi$). Hence, L is not $CAT(1)$.

Conversely, suppose that L is a flag complex. Then the link of each vertex is also a flag complex and by induction on dimension we may assume that $K(L) \leqslant 1$. Hence, it suffices to show every closed geodesic in L has length $\geqslant 2\pi$. Suppose, to the contrary, that α is a closed geodesic with $\ell(\alpha) < 2\pi$. Let L' be the full subcomplex of L spanned by the set of vertices v such that $\alpha \cap \mathrm{Star}(v) \neq \emptyset$. (Here $\mathrm{Star}(v)$ denotes the open star of v.) By Sublemma 1.25, α cannot intersect two disjoint open stars. Hence any two vertices of L' must be connected by an edge. Since L is a flag complex, this implies L' is an all right simplex. But this is impossible since a simplex contains no closed geodesic. □

(b) *Moussong's Lemma*

DEFINITION 1.26. A spherical simplex has *size* $\geqslant \pi/2$ if all of its edge lengths are $\geqslant \pi/2$.

DEFINITION 1.27. Let L be a PS simplicial complex with simplices of size $\geqslant \pi/2$. L is a *metric flag complex* if given a set of vertices $\{v_0, \ldots, v_k\}$, which are pairwise joined by edges, such that there exists a spherical k-simplex with these edge lengths, then $\{v_0, \ldots, v_k\}$ spans a k-simplex in L.

LEMMA 1.28 (Moussong's Lemma). *Let L be a PS simplicial complex with simplices of size $\geqslant \pi/2$. Then L is $CAT(1)$ if and only if it is a metric flag complex.*

This generalization of Gromov's Lemma is the main technical result in the Ph.D. thesis of Moussong [51]. Its proof is quite a bit more difficult than that of Gromov's Lemma and we will not try to explain it here. We will use it in the next chapter to show that a certain PE complex associated to any Coxeter group is $CAT(0)$.

DEFINITION 1.29. A cell is *simple* if the link of each vertex is a simplex.

The edge lengths of such a simplex are interior angles in the 2-dimensional faces. Thus, such a simplex has size $\geqslant \pi/2$ if all such angles in the 2-cells are $\geqslant \pi/2$.

COROLLARY 1.30. *Let X be a PE complex with simple cells and with 2-cells having nonacute angles. Then $K(X) \leqslant 0$ if and only if the link of each vertex is a metric flag complex.*

(c) *Orthogonal joins.* Suppose that $\sigma_1 \subset S^{k_1}$ and $\sigma_2 \subset S^{k_2}$ are spherical cells. Regard S^{k_1} and S^{k_2} as a pair of orthogonal great subspheres in $S^{k_1+k_2+1} \subset \mathbb{R}^{k_1+1} \times \mathbb{R}^{k_2+1}$. Then the *orthogonal join* $\sigma_1 * \sigma_2$ of σ_1 and σ_2 is the union of all geodesic segments from σ_1 to σ_2 in $S^{k_1+k_2+1}$. It is naturally a spherical cell of dimension equal to $\dim \sigma_1 + \dim \sigma_2 + 1$. If L_1 and L_2 are PS cell complexes, then their orthogonal join $L_1 * L_2$ is defined to be the union of all cells $\sigma_1 * \sigma_2$ where σ_1 is a cell in L_1 and σ_2 is a cell in L_2. It is naturally a PS cell complex, homeomorphic to the usual topological join of the underlying polyhedra.

The following result is proved in the appendix of [20].

PROPOSITION 1.31. *If L_1 and L_2 are $CAT(1)$, PS cell complexes, then $L_1 * L_2$ is $CAT(1)$.*

REMARK 1.32. For example, taking L_2 to be a point we see if L_1 is $CAT(1)$, then so is the "spherical cone" on L_1. Similarly, taking $L_2 = S^0$, we see that the "spherical suspension" of L_1 is $CAT(1)$.

(d) *Spherical buildings.* Tits has defined a certain remarkable class of simplicial complexes called "buildings", e.g., see [17] and [55]. Associated to a building B there is a Coxeter group W. (This will be defined in Section 2.) The building B can be written as a union of apartments A_α,

$$B = \bigcup A_\alpha,$$

where each A_α is isomorphic to the Coxeter complex for W. If W is a finite group, then this Coxeter complex can naturally be thought of as a triangulation of S^n, the round n-sphere, for some n. The building is called *spherical* if its associated Coxeter group is finite (so that each apartment is a round sphere). Thus, a spherical building has a natural structure of a PS simplicial complex.

The axioms for buildings imply that any two points in B lie in a common apartment. Furthermore, the geodesic between them also lies in this apartment. From this we can immediately deduce the following. (See also [32].)

THEOREM 1.33. *Any spherical building is $CAT(1)$.*

EXAMPLE 1.34. A *generalized m-gon* is a connected, bipartite graph of diameter m and girth $2m$. (A graph is *bipartite* if its vertices can be partitioned into two sets so that no two vertices in different sets span an edge. The *diameter* of a graph is the maximum distance between two vertices, its *girth* is the minimum length of a circuit.) A 1-dimensional spherical building is the same thing as a generalized m-gon ($m \neq \infty$). The piecewise spherical structure is defined by declaring each edge to have length π/m.

(e) *Polar duals of hyperbolic cells.* Suppose that C^n is a convex n-cell in hyperbolic n-space $\mathbb{H}^n$. Let F be a face of codimension k in C^n, $k \geqslant 1$. Choose a point x in the relative interior of F and consider the unit sphere S^{n-1} in the tangent space $T_x\mathbb{H}^n$. The set of outward-pointing unit normals to the codimension one faces which contain F span a spherical $(k-1)$-cell in S^{n-1} which we denote by σ_F. (Roughly, σ_F is the set of all outward pointing unit normals at F.) The *polar dual* of C^n is defined as

$$P(C^n) = \bigcup_F \sigma_F.$$

It is a PS cell complex, which, it is not difficult to see, is homeomorphic to S^{n-1}. For further details, see [23].

REMARK 1.35. (i) The same construction can be carried out in $\mathbb{E}^n$ or S^n. For a cell in $\mathbb{E}^n$ its polar dual is a PS cell complex which is isometric to the round $(n-1)$-sphere. For a cell C^n in S^n, its polar dual is just the boundary of the dual cell C^*, where $C^* = \{x \in S^n \mid d(x, C^n) \geqslant \pi/2\}$. In all three cases, the cell structure on $P(C^n)$ is combinatorially equivalent to the boundary to the dual polytope to C^n.

(ii) If we use the quadratic form model for $\mathbb{H}^n$, and $C^n \subset \mathbb{H}^n$, then $P(C^n)$ is naturally a subset of the unit pseudosphere, $S^n_1 = \{x \in \mathbb{R}^{n+1} \mid \langle x, x\rangle = 1\}$, where $\langle x, x\rangle = -(x_1)^2 + (x_2)^2 + \cdots + (x_{n+1})^2$.

THEOREM 1.36. *Suppose C^n is a convex cell in $\mathbb{H}^n$. Then $P(C)$ is $CAT(1)$.*

When $n = 2$, $P(C^2)$ is a circle, the length of which is the sum of the exterior angles of C^2. By the Gauss–Bonnet Theorem, this sum is $2\pi + \text{Area}(C^2)$. This completes the proof for $n = 2$. When $n = 3$, the theorem is due to Rivin and Hodgson [53]. For $n > 3$, it appears in [23].

FURTHER REMARKS 1.37. (i) A stronger result is actually true. The length of any closed geodesic in $P(C^n)$ is strictly greater than 2π. (As we saw, for $n = 2$, this follows from the Gauss–Bonnet Theorem.) Furthermore, the same is true for the link of every cell in $P(C^n)$ (since such a link is, in fact, the polar dual of some face of C^n). Sometimes I have defined a PS cell complex to be "large" if it is $CAT(1)$. Perhaps PS complexes satisfying the above stronger condition should be "extra large".

(ii) The definition of polar dual makes sense for any intersection of half-spaces in $\mathbb{H}^n$ (compact or not) and it is proved in [23] that these polar duals are also $CAT(1)$.

(iii) The main argument of [53] is in the converse direction. They show that any PS structure on S^2 which is extra large arises as the polar dual of a 3-cell in $\mathbb{H}^3$ (unique up to isometry). An analogous result, relating metrics with $K \geqslant 1$ on S^2 to convex surfaces in $\mathbb{E}^3$ had been proved much earlier by Aleksandrov.

(iv) My main interest in Theorem 1.36 is that it provides a method for constructing a large number of examples of $CAT(1)$, PS structures on S^{n-1}, which are not covered by Moussong's Lemma. Moreover, if we deform a convex cell in $\mathbb{H}^n$ we obtain a large family of deformations of its polar dual through $CAT(1)$ structures.

(v) A nice application of Theorem 1.36 is given in [26]. Identify $\mathbb{H}^n$ with the sheet of the hyperboloid in $\mathbb{R}^{n+1}$ defined by $\langle x, x\rangle = -1$ and $x_{n+1} > 0$. Let V be any discrete subset of $\mathbb{H}^n$ such that each Dirichlet domain for V is bounded. Take the convex hull of V in $\mathbb{R}^{n+1}$ and let $B(V)$ denote its boundary. It is not hard to see that the restriction of the bilinear form to the tangent space of any face in $B(V)$ is positive definite and hence, that $B(V)$ has a natural PE structure. Moreover, for any vertex v in V, the link of v in $B(V)$ can be identified with the polar dual of the Dirichlet cell centered at v. So, by Theorem 1.36, $B(V)$ is $CAT(0)$. As a corollary we have that any complete hyperbolic manifold has a nonpositively curved, PE structure.

(f) *Branched covers of round spheres.* Suppose that M^n is a smooth Riemannian manifold and that $p : \widetilde{M}^n \to M^n$ is a branched covering by some other manifold $\widetilde{M}^n$. The metric on M^n induces a (non-Riemannian) metric on $\widetilde{M}^n$.

QUESTION. If M^n has sectional curvature $\leqslant \varepsilon$, then when is $K(\widetilde{M}^n) \leqslant \varepsilon$?

We further suppose that the branching is locally modeled on $\mathbb{R}^n \to \mathbb{R}^n/G$ where G is some finite linear group. (Since M^n is a manifold we must therefore have that $\mathbb{R}^n/G$ is homeomorphic to $\mathbb{R}^n$.)

The following two conditions are easily seen to be necessary for $K(\widetilde{M}) \leqslant \varepsilon$:

(i) $K(M) \leqslant \varepsilon$,

(ii) locally, the closure of each stratum of the branched set is a convex subset of M.

Let $x \in M$ be a branch point and let S_x be the unit sphere in T_xM. There is an induced finite sheeted branched cover $\widetilde{S}_x \to S_x$. Since the branched set in S_x must satisfy (ii), it follows that the metric on $\widetilde{S}_x$ (induced from the round metric on S_x) is piecewise spherical. We think to $\widetilde{S}_x$ as the "link" at a point $\tilde{x} \in p^{-1}(x)$. It turns out ([20, Theorem 5.3]) that together with (i) and (ii) the following condition is necessary and sufficient for $K(\widetilde{M})$ to be $\leqslant \varepsilon$:

(iii) $\widetilde{S}_x$ is $CAT(1)$, for all branch points x.

Therefore, the answer to our question is closely tied to the question of when the branched cover of a round sphere is $CAT(1)$. A detailed study of this question is made in [20].

For example, suppose that G is a finite, noncyclic subgroup of $SO(3)$ (so that G is either dihedral or the group of orientation-preserving symmetries of a regular solid). Then S^2/G is homeomorphic to S^2 and $S^2 \to S^2/G$ has three branch points. Choose three points x_1, x_2 and x_3 in the round 2-sphere S^2 and assign x_i a branching order of m_i, where $\sum(1/m_i) > 1$. Let $\widetilde{S}^2 \to S^2$ be the corresponding $|G|$-fold branched cover. In [20] we prove the following result.

PROPOSITION 1.38. *$\widetilde{S}^2$ is $CAT(1)$ if and only if*

(i) *x_1, x_2, and x_3 lie on a great circle in S^2, but are not contained in any semi-circle, and*

(ii) *$d(x_i, x_j) \geqslant \pi/m_k$, where (i, j, k) is any permutation of $(1, 2, 3)$.*

1.4. *Infinitesimal shadows and the ideal boundary*

Suppose that X is a piecewise constant curvature polyhedron and that $c:(-a, a) \to X$ is a piecewise geodesic segment. Subdividing, we get two piecewise geodesic arcs $c_{\text{out}}:[0, a) \to X$ and $c_{\text{in}}:[0, a) \to X$, defined by $c_{\text{out}}(t) = c(t)$ and $c_{\text{in}}(t) = c(-t)$, and two unit tangent vectors c'_{out} and c'_{in} in $Lk(c(0), X)$. The proof of the following lemma is left as a straightforward exercise for the reader.

LEMMA 1.39. *A piecewise geodesic path $c:(-a, a) \to X$ is a local geodesic at $c(0)$ (i.e., its restriction to some smaller interval about 0 is a geodesic) if and only if the distance from c'_{in} to c'_{out} in $Lk(c(0), X)$ is $\geqslant \pi$.*

EXAMPLE 1.40. Let $S^1(2\pi + \delta)$ denote a circle of circumference $2\pi + \delta$, and let X denote the Euclidean cone on $S^1(2\pi + \varepsilon)$. Thus, $x = ([0, \infty) \times S^1(2\pi + \delta))/\sim$, where $(r, \theta) \sim (r', \theta')$ if and only if $r = r' = 0$. The metric is given by the usual formula for metric on $\mathbb{R}^2$

in polar coordinates (see [15]). Let $\theta_1, \theta_2 \in S^1(2\pi + \delta)$ be two points with $d(\theta_1, \theta_2) \geqslant \pi$ and consider the path $c: \mathbb{R} \to X$ defined by

$$c(t) = \begin{cases} (-t, \theta_1) & \text{for } t \leqslant 0, \\ (t, \theta_2) & \text{for } t \geqslant 0. \end{cases}$$

It follows from 1.39, that c is a geodesic in X. Thus, if $\delta > 0$, there are many possible ways to continue the geodesic ray $c|_{(-\infty,0]}$ to a geodesic line. These continuations are parametrized by an arc of θ_2's, in fact, by the arc of radius $\frac{1}{2}\delta$ about the point of distance $\pi + \frac{1}{2}\delta$ from θ_1.

This example illustrates a dramatic difference between singular metric spaces (e.g., piecewise constant curvature polyhedra) and Riemannian manifolds: in the singular case extensions of geodesic segments are not necessarily unique.

Suppose that X is a piecewise constant curvature polyhedron, that $x \in X$ and that $v \in Lk(v, X)$. The *infinitesimal shadow of x with respect to v*, denoted by $\mathrm{Shad}(x, v)$, is the subset of $Lk(x, X)$ consisting of all $w \in Lk(x, X)$ such that there is a geodesic $c: (-a, a) \to X$ with $c(0) = x$, $c'_{\text{in}} = v$ and $c'_{\text{out}} = w$. Thus, $\mathrm{Shad}(x, v)$ measures the possible outgoing directions of possible extensions for a geodesic coming into x from the direction v. In particular, X has "extendible geodesics" if and only if $\mathrm{Shad}(x, v)$ is nonempty for all choices of x and v.

For example, if $Lk(x, X)$ is isometric to the round sphere S^{n-1}, then it follows from Lemma 1.39 that the infinitesimal shadow of x with respect to v is the antipodal point $-v$.

As another example, suppose that X is the cone on $S^1(2\pi + \delta)$, $\delta \geqslant 0$, as in Example 1.40. Then any infinitesimal shadow at the cone point is an arc of radius $\frac{1}{2}\delta$ in $S^1(2\pi + \delta)$. Combining this example with Lemma 1.39, we get the following result.

LEMMA 1.41. *Suppose that X is a piecewise constant curvature polyhedron, that $x \in X$ and that $v \in Lk(x, X)$. Then*

$$\mathrm{Shad}(x, v) = Lk(x, X) - B_\pi(v),$$

where $B_\pi(v)$ denotes the open ball in $Lk(x, X)$ of radius π about v.

The ideal boundary. Now suppose that X is a proper $CAT(0)$ space. ("Proper" means that closed metric balls are compact.) Then X can be compactified to a space $\overline{X}$ by adding an "ideal boundary" $X(\infty)$. Here is the idea.

Fix a base point $x_0 \in X$. We compactify X by adding an endpoint $c(\infty)$ to each geodesic ray $c: [0, \infty) \to X$, with $c(0) = x_0$. Thus, $X(\infty)$ is the set of geodesic rays beginning at the base point.

The topology on $\overline{X}$ can be described as follows. Let $z = c(\infty) \in X(\infty)$ and let U be an open neighborhood of $c(r)$ in $S_{x_0}(r)$ (the sphere of radius r about x_0) and let V be the set of all points of the form $b(t)$, $t \in (r, \infty]$, where $b: [0, \infty) \to X$ is a geodesic ray starting at x_0 and passing through U. The sets V a neighborhood basis at z. At a point $x \in X$, a neighborhood basis consists of the open balls centered at x.

For $s > r$, by using geodesic contraction, one can define a natural projection

$$p: \overline{B}_{x_0}(s) \to \overline{B}_{x_0}(r)$$

from one closed ball to a smaller one. If $c: [0, s] \to X$ is a geodesic with $c(0) = x_0$, then p is defined by

$$p(c(t)) = \begin{cases} c(r) & \text{if } t \geqslant r, \\ c(t) & \text{if } t \leqslant r. \end{cases}$$

This gives inverse systems of maps $\overline{B}_{x_0}(s) \to \overline{B}_{x_0}(r)$ and $S_{x_0}(s) \to S_{x_0}(r)$. Clearly, $X(\infty) = \varprojlim S_{x_0}(r)$ and if X has extendible geodesics, then $\overline{X} = \varprojlim \overline{B}_{x_0}(r)$. Moreover, the topologies on $\overline{X}$ and on $X(\infty)$ are those of the inverse limits.

EXAMPLE 1.42. If X is a complete $CAT(0)$ Riemannian n-manifold, then a geodesic ray starting at x_0 is uniquely determined by its unit tangent vector at x_0. It follows that $\overline{X}$ is homeomorphic to the n-disk and $X(\infty)$ to its boundary S^{n-1}.

A serious problem with the above definition of the ideal boundary is that it seems to depend on the base point x_0. A definition can be given which is independent of the choice of base point. One way to do this is to define an equivalence relation, *asymptoty*, on the set of unbased geodesic rays. Two such rays are *asymptotic* if they remain a bounded distance apart. One can then prove that the natural map {rays based at x_0} $\to$ {asymptoty classes of rays} is a bijection. Thus, $X(\infty)$ is the set of asymptoty classes of rays.

A third description of $\overline{X}$ and $X(\infty)$ in terms of "horofunctions" is given in [6]. That these definitions coincide in the Riemannian case is proved in [6]. The fact that the arguments of [6] can be extended to the general case is explained in [35] and [38] as well as in [15].

1.5. *Some well-known results in geometric topology and their implications for $CAT(0)$ spaces*

A closed n-manifold is *homology sphere* if it has the same homology as does S^n. Is every homology n-sphere homeomorphic to S^n? Poincaré asked this question and came up with a counterexample for $n = 3$. His new homology sphere M^3 was S^3/G where S^3 is the group of quaternions of length one and where G is the binary icosahedral group (a subgroup of order 120). To distinguish M^3 from S^3 Poincaré discovered the concept of the fundamental group and noted that $\pi_1(M^3) = G$ while $\pi_1(S^3) = 1$. Later it was shown that for each $n \geqslant 3$ there exist homology n-spheres which are not simply connected. In fact, Kervaire showed in [48] that for $n \geqslant 5$ the fundamental group could be any finitely presented group G satisfying $H_1(G) = H_2(G) = 0$. On the other hand, the Generalized Poincaré Conjecture asserts that any simply connected homology n-sphere, $n > 1$, is homeomorphic to S^n. For $n \geqslant 5$, this was proved by Smale, in the smooth case, (and then generalized to the PL case by Stallings and the topological case by Newman); for $n = 4$, it was proved by Freedman in [42].

Suppose C^n is a compact contractible n-manifold with boundary. Is C^n homeomorphic to the n-disk? Poincaré duality implies that ∂C^n is a homology $(n-1)$-sphere, but there is no reason for it to be simply connected. In fact, one can prove that any homology $(n-1)$-sphere can be realized as the boundary of a contractible n-manifold. (The most difficult case, $n = 4$, was proved in [42].) Hence, it follows from the previous paragraph that, for $n \geqslant 4$, there are examples where C^n is not homeomorphic to the n-disk. (The 3-dimensional Poincaré Conjecture asserts that there are no such examples for $n = 3$.) On the other hand, for $n \geqslant 5$, the Generalized Poincaré Conjecture implies that C^n is homeomorphic to the n-disk if and only if ∂C^n is simply connected.

Similarly, we can ask to what extent do open contractible n-manifolds differ from the interior of D^n (i.e., from $\mathbb{R}^n$). If C^n is a compact contractible manifold, $n > 2$, and if ∂C^n is not simply connected, then the interior of C^n is not homeomorphic to $\mathbb{R}^n$. The reason is that the interior of C^n is not simply connected at infinity. (Its "fundamental group at infinity" is isomorphic to $\pi_1(\partial C^n)$.) Thus, for $n \geqslant 4$ there are contractible manifolds (without boundary) which are not homeomorphic to $\mathbb{R}^n$. This is also true for $n = 3$: there is a famous example of J.H.C. Whitehead of a contractible 3-manifold which is not simply connected at infinity. In fact, the situation is more complicated than in the compact case: the end of an open contractible manifold need not be "tame", i.e., the contractible manifold might not be homeomorphic to the interior of any compact manifold with boundary. For example, the fundamental group at infinity need not be finitely generated. On the other hand, Stallings [59] (for $n \geqslant 5$) and Freedman [42] (for $n = 4$) showed that a contractible n-manifold is homeomorphic to $\mathbb{R}^n$ if and only if it is simply connected at infinity.

Here are a few questions which we shall be concerned with.

QUESTION 1.43. If a contractible n-manifold admits a cocompact transformation group, then is it homeomorphic to $\mathbb{R}^n$? In particular, if M^n is a closed aspherical manifold, then is its universal cover homeomorphic to $\mathbb{R}^n$?

By constructing examples from reflection groups, we shall see in Section 2.5 that the answer is "no" for each $n \geqslant 4$. This reflection group construction suggests the following question.

QUESTION 1.44. Suppose M^n, $n > 2$, is a closed aspherical n-manifold with universal cover $\widetilde{M}^n$. If the fundamental group at infinity of $\widetilde{M}^n$ is finitely generated, then is it homeomorphic to $\mathbb{R}^n$? (The *fundamental group at infinity* $\pi_1^\infty(\widetilde{M}^n)$ is the inverse limit $\varprojlim \pi_1(\widetilde{M}^n - K)$ where K ranges over the compact subsets of $\widetilde{M}^n$.)

Something very close to an affirmative answer to this has been proved by Wright [68]. He shows that the inverse system $\pi_1(\widetilde{M}^n - K)$ cannot be "pro-monomorphic".

Next we turn to some questions involving nonpositively curved spaces.

QUESTION 1.45. Suppose X^n is a $CAT(0)$ manifold.

(a) (Gromov) Is X^n homeomorphic to $\mathbb{R}^n$?
(b) If so, is $(\overline{X}, X(\infty))$ homeomorphic to (D^n, S^{n-1})? In particular, if X^n is simply connected at infinity, then is $X(\infty)$ homeomorphic to S^{n-1}?
(c) If $X(\infty)$ is a manifold, then is it homeomorphic to S^{n-1}?

The answers to all the questions in 1.45 are "no" for $n > 4$ and "yes" for $n \leqslant 3$. Paul Thurston [64] has proved that the answer to (a) is also "yes" when $n = 4$, provided the manifold has a least one "tame point," but without this hypothesis the answer is not known. The result for $n = 3$ is essentially due to Rolfsen [54]. For a clear discussion of these questions in dimensions 3 and 4, the reader is referred to [64].

QUESTION 1.46. If X^n is the universal cover of a nonpositively curved, closed manifold, then are the answers to Questions 1.45 the same?

A polyhedron N^n is a PL *n-manifold* if for each k-cell σ in N, $Lk(\sigma, N^n)$ is piecewise linearly homeomorphic to S^{n-k-1} (where a sphere has a standard PL structure as the boundary of a convex cell).

The next result, which is proved in [35], asserts that in the PL context, Questions 1.45 have affirmative answers.

THEOREM 1.47 (Stone [62] and [35]). *Suppose X^n is a piecewise Euclidean or piecewise hyperbolic polyhedron and that*

(a) *X^n is a PL manifold (i.e., the underlying polyhedral structure on X^n is that of a PL manifold), and*

(b) *X^n is $CAT(0)$.*

Then $(\overline{X}, X(\infty))$ is homeomorphic to (D^n, S^{n-1}).

The result of [62] states that X^n is PL homeomorphic to $\mathbb{R}^n$.

The proof of 1.47 uses the Approximation Theorem for cell-like maps due to Siebenmann ($n \geqslant 5$), Quinn ($n = 4$), Armentrout ($n = 3$), and Moore ($n = 2$). Before stating this theorem, we recall some definitions. (Our discussion is taken from [39].)

A compact metrizable space C is *cell-like* if there is an embedding of C into the Hilbert cube I^∞ so that for any neighborhood U of C in I^∞, the space C is contractible in U. A cell-like subspace of a manifold is *cellular* if it has arbitrarily small neighborhoods homeomorphic to a cell. A compact subset of S^n is *pointlike* if its complement is homeomorphic to $\mathbb{R}^n$. A pointlike subset of S^n is cellular in S^n. A map is *cell-like* if each point inverse image is cell-like.

THEOREM 1.48 (The Approximation Theorem). *Suppose $f: M^n \to N^n$ is a cell-like map of topological manifolds. If $n = 3$, further assume that f is cellular (i.e., each point inverse image is cellular). Then f can be approximated by a homeomorphism.*

A map which can be approximated by a homeomorphism is a *near homeomorphism*. We shall also need the following theorem of Brown [18].

THEOREM 1.49 (M. Brown). *An inverse limit of near homeomorphisms is a near homeomorphism.*

SKETCH OF PROOF OF THEOREM 1.47. The rough idea is that one shows that the hypotheses of Theorem 1.47 imply that each infinitesimal shadow is cell-like. Since these shadows are basically point inverse images of the geodesic contraction map, one can then conclude from the Approximation Theorem that each closed metric ball is homeomorphic to an n-disk and from Brown's Theorem that

$$(\overline{X}, X(\infty)) \cong (D^n, S^{n-1}).$$

To be more explicit, let (T_n) denote the statement of the theorem in dimension n and let (L_n) denote the following statement:

If M^n has a PS structure such that (a) M^n is a PL manifold and (b) M^n is $CAT(1)$, then for each $r \in (0, \pi)$ and $v \in M^n$ the closed metric ball $\overline{B}_v(r)$ is homeomorphic to D^n and $B_v(\pi)$ is homeomorphic to $\mathbb{R}^n$.

The inductive scheme is then $(L_{n-1}) \Rightarrow (L_n)$ and (T_n). For example, to see that $(L_{n-1}) \Rightarrow (T_n)$, let $x \in X^n$ and $v \in Lk(x, X)$. By hypothesis, $Lk(x, X) \cong S^{n-1}$ and by (L_{n-1}), $B_v(\pi) \cong \mathbb{R}^{n-1}$. Hence, by Lemma 1.43, $\text{Shad}(x, v) = Lk(x, X) - B_v(\pi)$ is pointlike. In particular, since $\text{Shad}(x, v)$ is nonempty, it follows that geodesics in X are extendible and from this that

$$(\overline{X}, X(\infty)) \cong \varprojlim(\overline{B}_{x_0}(r), \partial\overline{B}_{x_0}(r))$$

for any base point x_0. Moreover, using the fact that infinitesimal shadows are cell-like it is easy to produce a cell-like map from $\overline{B}_{x_0}(r)$ to the disk of radius r in $\mathbb{R}^n$. Further details can be found in Section 3 of [35]. □

A space N is a *homology n-manifold* if for each $x \in N$, $H_i(N, N - x)$ vanishes for $i \neq n$ and is infinite cyclic for $i = n$. If N is a polyhedron, then it is a homology n-manifold if and only if for each k-cell σ of N, $Lk(\sigma, N)$ has the same homology as does S^{n-k-1}.

THEOREM 1.50. *Suppose X^n is a piecewise Euclidean or piecewise hyperbolic polyhedron and that*

(a) *X^n is a homology n-manifold, and*

(b) *X^n is $CAT(0)$.*

Then X has extendible geodesics and its ideal boundary $X(\infty)$ is a homology $(n-1)$-manifold with the same homology as S^{n-1}.

The point is that the arguments in the proof of Theorem 1.47 show that under the hypotheses of Theorem 1.50, each infinitesimal shadow is acyclic.

REMARK 1.51. The proofs of Theorems 1.47 and 1.50 are indicative of a (conjectural) "local-to-global" principle: the topology of the ideal boundary of a $CAT(0)$ space should be controlled by the topology of the infinitesimal shadows (reflecting the topology of the links of points). For example in Theorem 1.47, the fact that the link of each vertex of X is a PL $(n-1)$-sphere is reflected in the fact that $X(\infty)$ is homeomorphic to S^{n-1}. In Theorem 1.50, the fact that the link of each vertex is a homology manifold with the same homology as S^{n-1} is reflected in the fact that $X(\infty)$ has the same properties.

Here is another result ([35, Proposition 3d.3, p. 374]) which clarifies the picture.

THEOREM 1.52. *Suppose that X is a CAT(0) polyhedron such that the link of each vertex is a PL manifold (i.e., X has isolated PL singularities). Then any metric sphere in X, which does not pass through a vertex, is homeomorphic to the connected sum of the links of the vertices which it encloses and $X(\infty)$ is the inverse limit of this sequence of increasing connected sums. In particular, if X is a polyhedral homology manifold (so that the link of each vertex is a PL manifold and a homology sphere), then, generically, each metric sphere is a connected sum of PL homology spheres and $X(\infty)$ is the resulting inverse limit.*

If a polyhedron is a topological manifold, then, of course, it is a homology manifold. Moreover, it can happen that the link of a k-cell, $k > 0$, in a polyhedral topological manifold can be a nonsimply connected homology sphere. This is the content of the famous Double Suspension Theorem (due to Edwards in many cases and to Cannon [19] in complete generality). This result states that if N^n is a PL manifold and a homology sphere, then its double suspension $S^1 * N^n$ is a topological manifold (which must be homeomorphic to S^{n+2} by the Generalized Poincaré Conjecture). The definitive result along these lines in the following Polyhedral Manifold Characterization Theorem of [39].

THEOREM 1.53 (Edwards). *A polyhedral homology manifold of dimension $\geqslant 5$ is a topological manifold if and only if the link of each vertex is simply connected.*

Thus, in spite of Theorem 1.47, the possibility remains that the answers to Questions 1.45 are negative. As we shall see in Section 2.5, this is, in fact, the case (the answers to 1.45(a), (b) and (c) are negative).

1.6. *Euler characteristics and the Combinatorial Gauss–Bonnet Theorem*

HOPF'S CONJECTURE. Suppose M^{2n} is a closed Riemannian manifold, with $K(M) \leqslant 0$. Then $(-1)^n\chi(M^{2n}) \geqslant 0$. (Here χ denotes the Euler characteristic.)

REMARK 1.54. (i) There is an analogous version of this conjecture for nonnegative curvature: the Euler characteristic should be nonnegative.

(ii) Thurston has conjectured that Hopf's Conjecture should hold for any closed, aspherical $2n$-manifold.

The reason for believing this is the Gauss–Bonnet Theorem (proved by Chern in dimensions > 2). Recall that this is the following theorem.

GAUSS–BONNET THEOREM.

$$\chi(M^{2n}) = \int P.$$

Here P is a certain $2n$-form called the "Pfaffian" or the "Euler form". This leads to the following.

QUESTION 1.55. Does $K(M^{2n}) \leqslant 0$ imply that $(-1)^n P \geqslant 0$? (In other words, is $(-1)^n P$ equal to the volume form multiplied by a nonnegative function?)

In dimension 2 the answer is, of course, yes, since P is then just the volume form times the curvature. The answer is also yes in dimension 4. A proof is given by Chern in [28], where the result is attributed to Milnor.

Hopf's Conjecture holds in higher dimensions under the hypothesis that the "curvature operator" is negative semi-definite (which is stronger than assuming that the sectional curvature is nonpositive). On the other hand, in dimensions $\geqslant 6$, Geroch [43] showed in 1976 that the answer to Question 1.55 is no.

The following combinatorial version of the Gauss–Bonnet Theorem is a classical result. A proof can be found in [27], where one can also find a convincing argument that it is the correct analog of the smooth Gauss–Bonnet Theorem.

THEOREM 1.56. *Suppose X is a finite, PE cell complex. Then*

$$\chi(X) = \sum_{v} P\big(Lk(v, X)\big).$$

Here P is a certain function which assigns a real number to any finite, PS cell complex. We define it below.

Let $\sigma \subset S^k$ be a spherical k-cell. Its dual cell σ^* is defined by $\sigma^* = \{x \in S^k \mid d(x, \sigma) \geqslant \pi/2\}$. Let $a^*(\sigma)$ be the volume of σ^* normalized so that volume of S^k is 1, i.e.,

$$a^*(\sigma) = \frac{\mathrm{vol}(\sigma^*)}{\mathrm{vol}(S^k)}.$$

If L is a finite, PS cell complex then $P(L)$ is defined by

$$P(L) = 1 + \sum_{\sigma} (-1)^{\dim \sigma + 1} a^*(\sigma),$$

where the summation is over all cells σ in L.

EXAMPLE 1.57. Suppose that σ is an all right k-simplex. Then σ^* is also an all right k-simplex. Since S^k is tessellated by 2^{k+1} copies of σ^* we see that $a^*(\sigma) = (\frac{1}{2})^{k+1}$. Now let L be an all right PS simplicial complex and f_i the number of i-simplices in L. Then

$$P(L) = 1 + \sum_{\sigma} \left(-\frac{1}{2}\right)^{\dim \sigma + 1} = 1 + \sum_{i} \left(-\frac{1}{2}\right)^{i+1} f_i.$$

The following conjecture asserts that the answer to the combinatorial version of Question 1.55 should always be yes.

CONJECTURE 1.58. Suppose that L^{2n-1} is a PS cell complex homeomorphic to S^{2n-1}. If L^{2n-1} is $CAT(1)$, then $(-1)^n P(L^{2n-1}) \geqslant 0$.

Thus, this conjecture implies Hopf's Conjecture for PE manifolds.

If L is a flag complex, then, by Gromov's Lemma and Example 1.57, we have the following special case.

CONJECTURE 1.59. Suppose that L^{2n-1} is a flag complex which triangulates S^{2n-1}. Then

$$(-1)^n\left(1+\sum\left(-\frac{1}{2}\right)^{i+1} f_i\right) \geqslant 0.$$

So, this conjecture implies Hopf's Conjecture for PE cubical complexes which are closed manifolds.

REMARK 1.60. Conjecture 1.59 is analogous to the Lower Bound Theorem in the combinatorics (a result concerning inequalities among the f_i for simplicial polytopes). For example, the Lower Bound Theorem of [67] asserts that for any simplicial complex L which triangulates S^3, we have $f_1 \geqslant 4f_0 - 10$. Conjecture 1.59 asserts that, if, in addition, L is a flag complex, then $f_1 \geqslant 5f_0 - 16$. Some evidence for these conjectures is provided by the following two propositions. The first result follows from recent work of Stanley [61] as was observed by Eric Babson.

PROPOSITION 1.61. *Suppose that L^{2n-1} is the barycentric subdivision of the boundary complex of a convex $2n$-cell (so that L is a flag complex). Then Conjecture* 1.59 *holds for L.*

PROPOSITION 1.62. *Suppose that L^{2n-1} is the polar dual of a convex cell C^{2n} in $\mathbb{H}^n$. Then Conjecture* 1.58 *holds for L.*

This proposition follows from a formula of Hopf (predating the general Gauss–Bonnet Theorem) which asserts the $(-1)^n P(L^{2n-1})$ is one-half the hyperbolic volume of C^{2n} (suitably normalized).

Further details about these conjectures and further evidence for them can be found in [22].

REMARK 1.63. A natural reaction to Conjecture 1.58 is that it might contradict Geroch's result. One could try to obtain such a contradiction as follows. Take a smooth Riemannian manifold M^{2n} whose curvature operator at a point x is as in Geroch's result. Then try to approximate M^{2n} near x by a PE cell complex with nonpositive curvature. By the main result of [27] the numbers $P(L)$ for L a link in the complex, should approximate the Pfaffian at x and hence, have the wrong sign. However, it is not clear that such a approximation exists. Thus, we are led to ask the following.

QUESTION 1.64. Suppose M is a Riemannian manifold with $K(M) \leqslant 0$ (we could even assume the inequality is strict). Is M homeomorphic to a PE cell complex X with $K(X) \leqslant 0$?

As we explained in Remark 1.37(v), the answer is yes in the constant curvature case. For our conjectures to be correct, the answer, in general, should be no.

2. Coxeter groups

Coxeter groups and Coxeter systems are defined in Section 2.1. Associated to a Coxeter system there is a simplicial complex called its "nerve". The basic result of Section 2.1 is Lemma 2.6, which asserts that any finite polyhedron can occur as the nerve of some Coxeter system. Eventually, this will be used to show that Coxeter groups provide a rich and flexible source of examples.

In Sections 2.2 and 2.3 we discuss a beautiful, PE cell complex Σ which is naturally associated to a Coxeter system (W, S). From the results of Section 1, we get the important result of Moussong (generalizing an earlier observation of Gromov), that Σ is nonpositively curved and hence, contractible (since it is simply connected). The connection with reflection groups is explained in Section 2.4.

In Section 2.5 we briefly discuss some important special cases of this construction.

2.1. *Coxeter systems*

DEFINITION 2.1. Let S be a finite set. A *Coxeter matrix* $M = (m_{ss'})$ is an $S \times S$ symmetric matrix with entries in $\mathbb{N} \cup \{\infty\}$ such that

$$m_{ss'} = \begin{cases} 1 & \text{if } s = s', \\ \geqslant 2 & \text{if } s \neq s'. \end{cases}$$

DEFINITION 2.2. Given a Coxeter matrix M, define a group W with presentation:

$$W = \langle S \mid (ss')^{m_{ss'}} = 1,\ \forall (s, s') \in S \times S \rangle.$$

W is called a *Coxeter group*.

If all the off-diagonal entries of M are 2 or ∞, then W is called *right-angled*.

Coxeter groups are intimately connected to the theory of reflection groups. This connection is not emphasized in this paper. For now it should suffice to mention that if a group W acts properly on a connected manifold and if W is generated by reflections (where a *reflection* is an involution whose fixed point set separates the manifold), then W is a Coxeter group (cf. [29, Theorem 4.1]).

Given M, it is proved in [13, Chapter V, §4.3, pp. 91–92] that one can find a family $(\rho_s)_{s \in S}$ of linear reflections $\rho_s : \mathbb{R}^S \to \mathbb{R}^S$ so that $\rho_s \circ \rho_{s'}$ has order $m_{ss'}$ for all $(s, s') \in S \times S$. It follows that the map $s \to \rho_s$ extends to a representation $\rho : W \to GL(\mathbb{R}^S)$ called the *canonical representation*. The existence of this representation immediately implies the following:

(a) the natural map $S \to W$ is an injection (and henceforth, we shall identify S with its image in W),

(b) order $(s) = 2$, for all $s \in S$,
(c) order $(ss') = m_{ss'}$, for all $(s, s') \in S \times S$.
The pair (W, S) is called a *Coxeter system.*

REMARK 2.3. It is proved in [13, Chapter V, §4.4, pp. 92–94] that the dual representation $\rho^* : W \to G((\mathbb{R}^S)^*)$ is faithful and has discrete image. Moreover, as explained in Section 3.1, W acts properly on certain open convex subset of $(\mathbb{R}^S)^*$. (These results are due to Tits.)

If T is a subset of S, then let W_T denote the subgroup of W generated by T.

LEMMA 2.4 ([13, p. 20]). *For any $T \subset S$, the pair (W_T, T) is a Coxeter system (i.e., its Coxeter matrix is $M|_T$).*

Let (W, S) be a Coxeter system. We define a poset, denoted $\mathcal{S}^f(W, S)$ (or simply $\mathcal{S}^f$) by

$$\mathcal{S}^f = \{T \mid T \subset S \text{ and } W_T \text{ is finite}\}.$$

It is partially ordered by inclusion. Consider $\mathcal{S}^f - \{\emptyset\}$. It is isomorphic to the poset of simplices of an abstract simplicial complex which we shall denote by $N(W, S)$ (or simply N). N is called the *nerve* of (W, S).

In other words, the vertex set of N is S and a subset T of S spans a simplex if and only if W_T is finite.

EXAMPLE 2.5. If W is finite, then N is the simplex on S.

Which finite polyhedra occur as the nerve of some Coxeter system? The next two results show that they all do. (Compare with Corollary 1.22.)

LEMMA 2.6. *Let L be any flag complex. Then there is a right-angled Coxeter system (W, S) with $N(W, S) = L$.*

PROOF. Let S be the vertex set of L and define a Coxeter matrix $(m_{ss'})$ by

$$m_{ss'} = \begin{cases} 1 & \text{if } s = s', \\ 2 & \text{if } \{s, s'\} \text{ spans an edge in } L, \\ \infty & \text{otherwise.} \end{cases}$$

If W is the associated right-angled Coxeter group, then $N(W, S) = L$. □

In particular, since the barycentric subdivision of any (regular) cell complex is a flag complex, we have the following corollary.

COROLLARY 2.7. *For any finite polyhedron P, there is a right-angled Coxeter system (W, S) with $N(W, S)$ homeomorphic to P.*

The main result of this section is the following theorem.

THEOREM 2.8 (Gromov, Moussong). *Associated to a Coxeter system (W, S) there is a PE cell complex $\Sigma(W, S)$ $(= \Sigma)$ with the following properties.*

(i) *The poset of cells in Σ is the poset of cosets*

$$WS^f = \coprod_{T \in S^f} W/W_T.$$

(ii) *W acts by isometries on Σ with finite stabilizers and with compact quotient.*

(iii) *Each cell in Σ is simple (so that for each vertex v, $Lk(v, \Sigma)$ is a simplicial cell complex). In fact, this complex is just $N(W, S)$.*

(iv) *Σ is $CAT(0)$.*

2.2. *Coxeter cells*

Throughout this subsection we suppose that W *is finite*. In this case, we will show that Σ can be identified as a convex cell in $\mathbb{R}^n$ $(n = \mathrm{Card}(S))$.

The canonical representation shows that W can be represented as an orthogonal linear reflection group on $\mathbb{R}^n$. The hyperplanes of reflection divide $\mathbb{R}^n$ into "chambers", each of which is a simplicial cone. (See in [13, p. 85].)

Choose a point x in the interior of some chamber. Define Σ to be the convex hull of Wx (the orbit of x). Σ is called a *Coxeter cell* of *type* W.

The proof of the next lemma is an easy exercise.

LEMMA 2.9. *Suppose W is finite.*

(i) *The vertex set of the Coxeter cell Σ is Wx.*

(ii) *Each face F of Σ is the convex hull of a set of vertices of the form $(wW_T)x$ for some $T \subset S$ and some coset wW_T of W_T. (So, F is a Coxeter cell of type W_T.)*

(iii) *The poset of faces of Σ is therefore,*

$$\coprod_{T \subset S} W/W_T.$$

(iv) *Σ is simple cell. $Lk(x, \Sigma)$ is the spherical $(n-1)$-simplex spanned by the outward pointing unit normals to the supporting hyperplanes of a chamber.*

REMARK 2.10. If x lies in a chamber with supporting hyperplanes indexed by S, then the vertex set of $Lk(x, \Sigma)$ is naturally identified with S. Moreover, the length of the edge from s to s' is $\pi - \pi/m_{ss'}$. (In other words, the corresponding angle in a 2-cell in Σ is $\pi - \pi/m_{ss'}$.)

EXAMPLE 2.11.

(i) If $W = \mathbb{Z}/2$, then Σ is an interval.

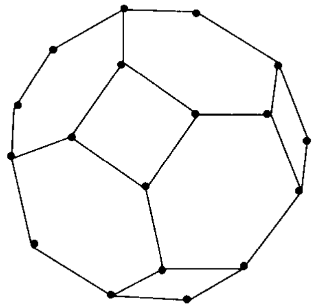


Fig. 4.

(ii) If $W = D_m$ (the dihedral group of order $2m$), then Σ is a $2m$-gon.
(iii) If (W, S) is the direct product of two Coxeter systems (W_1, S_1) and (W_2, S_2) (so that $W = W_1 \times W_2$ and $S = S_1 \coprod S_2$), then $\Sigma(W, S) = \Sigma(W_1, S_1) \times \Sigma(W_2, S_2)$. In particular, if $W = (\mathbb{Z}/2)^n$, then Σ is a n-dimensional box (= the product of n intervals).
(iv) If $W = S_n$, the symmetric group on n letters, then Σ is the $(n-1)$-cell called the *permutahedron*. The picture for $n = 4$ is given in Figure 4.

REMARK 2.12. By choosing x to be of distance 1 from each supporting hyperplane we can normalize each Coxeter cell so that every edge length is 2.

2.3. *The cell complex Σ (in the case where W is infinite)*

There is an obvious way to generalize the material of the previous subsection to the case where W is infinite. The cell complex Σ is defined as follows. The vertex set of Σ is W. Take a Coxeter cell of type W_T for each coset wW_T, $T \in \mathcal{S}^f$. Identify the vertices of this Coxeter cell with the elements of wW_T. Identify two faces of two Coxeter cells if they have the same set of vertices. This completes the definition of Σ as a cell complex.

If we normalize each Coxeter cell as in Remark 2.12, then the faces of the cells are identified isometrically and hence, Σ has the structure of a PE cell complex.

REMARK 2.13. Let $\lambda : S \to (0, \infty)$ be a function. If, in the definition of each Coxeter cell, we choose the point x to be of distance $\lambda(s)$ from the hyperplane corresponding to s, then the Coxeter cells again fit together to give a PE structure on Σ. We have arbitrarily chosen λ to be the constant function.

REMARK 2.14. By construction, the poset

$$W\mathcal{S}^f = \coprod_{T \in \mathcal{S}^f} W/W_T$$

is the poset of cells in Σ. If $\mathcal{P}$ is any poset, then let $\mathcal{P}'$ be its derived complex defined as in Example 1.19 (i.e., $\mathcal{P}'$ is the poset of finite chains in $\mathcal{P}$). $\mathcal{P}'$ is the poset of simplices in an abstract simplicial complex. Moreover, if $\mathcal{P}$ is the poset of cells in a cell complex, then $\mathcal{P}'$ is the poset of simplices in its barycentric subdivision. Applying these remarks to the case at hand, we see that the barycentric subdivision Σ' of Σ is just the geometric realization of $(W\mathcal{S}^f)'$. Alternatively, we could have defined Σ' as the geometric realization of $(W\mathcal{S}^f)'$ and then remarked that the poset of chains which terminate in wW_T can naturally be identified with the set of simplices in the barycentric subdivision of a Coxeter cell of type W_T. Hence, the cellulation of Σ by Coxeter cells could be recovered from Σ' by collecting together the appropriate simplices.

The link of a vertex. The group W acts isometrically on Σ and freely and transitively on its vertex set. Thus, there is an isometry of Σ which takes any vertex onto the element $1 \in W$. What is $Lk(1, \Sigma)$ (as a simplicial complex)? A cell contains the element 1 if and only if it corresponds to some identity coset W_T. Hence, the poset of simplices in $Lk(1, \Sigma)$ is just $\mathcal{S}^f - \{\emptyset\}$, i.e.,

$$Lk(1, \Sigma) = N(W, S).$$

What is the induced PS structures on N? Two distinct vertices s and s' of N are connected by an edge $e_{ss'}$ if and only if $m_{ss'} \neq \infty$. By Remark 2.10, $\ell(e_{ss'}) = \pi - \pi/m_{ss'}$ (where ℓ stands for length). Since a spherical simplex is determined by its edge lengths this determines PS structure on N.

A proof of the following lemma can be found in [13, p. 98].

LEMMA 2.15. *Let T be a subset of S. Consider the $T \times T$ matrix, $(c_{ss'})$, where $s, s' \in T$, and where $c_{ss'} = \cos(\pi - \pi/m_{ss'})$. Then W_T is finite if any only if $(c_{ss'})$ is positive definite.*

PROOF. Consider the canonical representation of W_T into $GL(n)$. Suppose W_T is finite. Then we may assume that the image of this representation is contained in $O(n)$. For each $s \in T$, let u_s be the outward-pointing unit normal to the hyperplane corresponding to s. Then $(u_s \cdot u_{s'}) = (c_{ss'})$ and hence, this matrix is positive definite since $(u_s)_{s\in T}$ is a basis for $\mathbb{R}^n$. Conversely, suppose that $(c_{ss'})$ is positive definite. Since the canonical representation preserves the corresponding bilinear form, we get that the image of this representation is contained in $O(n)$. Since this representation is also discrete and faithful (cf. Remark 2.3), W_T is a discrete subgroup of $O(n)$; hence, finite. □

COROLLARY 2.16. *$N(W, S)$ is a metric flag complex.*

We can now prove the main result.

PROOF OF THEOREM 2.8. We have already demonstrated the required properties of Σ except for (iv). Thus, it suffices to prove that Σ is $CAT(0)$. First of all, it is easy to see that Σ is simply connected. (One argument is to observe that the 2-skeleton of Σ is just the universal cover of the 2-complex associated to standard presentation of W.) Hence, by

Theorem 1.5, it suffices to show $K(\Sigma) \leqslant 0$. The link of any vertex is isometric to N. By Corollary 2.16 and Moussong's Lemma (Lemma 1.28) N is $CAT(1)$. Therefore, $K(\Sigma) \leqslant 0$ (by Theorem 1.11). □

REMARK 2.17. Theorem 2.8 was proved by Gromov [45, pp. 131–132] in the special case where W is right-angled. The general case was proved in [51]. The point is that in the right-angled case Σ is a cubical complex so we can use Gromov's Lemma (Lemma 1.21) rather than Moussong's generalization of it.

Combining Lemma 2.6 and Theorem 2.8 (in the right-angled case) we get the following.

COROLLARY 2.18. *Given a finite flag complex L, there is a finite, nonpositively curved, cubical PE complex X such that the link of each vertex in X is isomorphic to L.*

PROOF. Let (W, S) be the right-angled Coxeter system associated to L and let Γ be any torsion-free subgroup of finite index in W. (For example, Γ could be the kernel of the obvious epimorphism $W \to (\mathbb{Z}/2)^S$.) Set $X = \Sigma(W, S)/\Gamma$. □

There is a simpler version of the above construction, which does not directly use Coxeter groups, and which gives the following slightly more general result.

PROPOSITION 2.19. *Let L be a finite simplicial complex. Then there is a finite, cubical PE complex X such that the link of each vertex in X is isomorphic to L. (Of course, by Lemma 1.21, L is nonpositively curved if and only if L is a flag complex.)*

PROOF. (Independently due to Babson and Bridson and Haefliger, [15].) Let S be the vertex set of L. Consider the cube $[-1, 1]^S$ in the Euclidean space $\mathbb{R}^S$ with standard basis $(e_s)_{s\in S}$. Let X be the cubical subcomplex of $[-1, 1]^S$ consisting of all faces parallel to a subspace of the form $\mathbb{R}^T$, where T is the vertex set of a simplex in L and where $\mathbb{R}^T$ is the subspace spanned by $\{e_s\}_{s\in T}$. The vertex set of X is $\{\pm 1\}^S$ and the link of each such vertex is naturally identified with L. □

2.4. *Reflection groups*

Suppose that (W, S) is a Coxeter system, that X is a space and that $(X_s)_{s\in S}$ is a family of closed subspaces. For each $x \in X$, let $S(x) = \{s \in S \mid x \in X_s\}$. Define an equivalence relation $\sim$ on $W \times X$ by: $(w, x) \sim (w', x')$ if and only if $x = x'$ and $w^{-1}w' \in W_{S(x)}$. Let $\mathcal{U}(W, X)$ denote the quotient space $(W \times X)/\sim$. The group W acts on $\mathcal{U}(W, X)$; the orbit space is X. Moreover, we can identify X with the image of $1 \times X$ in $\mathcal{U}(W, X)$. Thus, X is a fundamental domain for the W-action. Any translate wX of X is called a *chamber*; thus, $\mathcal{U}(W, X)$ is decomposed into chambers. Each s in S acts on $\mathcal{U}(W, X)$ as a "reflection" in the following sense: the fixed point set of s separates $\mathcal{U}(W, X)$ into two "half-spaces" which are interchanged by s. (A more detailed discussion and further properties of this construction can be found in [66] or in [29].)

If Y is a space with an action of a Coxeter group W, then W is called a *reflection group* on Y if Y is equivariantly homeomorphic to $\mathcal{U}(W, X)$ for some subspace X of Y.

EXAMPLE 2.20. Suppose that X is the cone on S and that X_s denotes the point s. Then $\mathcal{U}(W, X)$ is the Cayley graph of (W, S).

EXAMPLE 2.21. Suppose that X is the geometric realization of the poset $\mathcal{S}^f$ of Section 2.1. Thus, a k-simplex in X corresponds to a chain $T_0 < T_1 < \cdots < T_k$, where $T_i \in \mathcal{S}^f$. For each s in S, let X_s denote the subcomplex consisting of those simplices such that $T_0 = \{s\}$. Then $\mathcal{U}(W, X)$ can be identified with the barycentric subdivision of the geometric realization $\Sigma(W, S)$ of the poset of cosets $W\mathcal{S}^f$.

Here is what is going on in the above example. Given any poset $\mathcal{P}$, let $|\mathcal{P}|$ denote the geometric realization of its derived complex. For each $p \in \mathcal{P}$ define subposets,

$$\mathcal{P}_{\leqslant p} = \{q \in \mathcal{P} \mid q \leqslant p\} \quad \text{and} \quad \mathcal{P}_{\geqslant p} = \{q \in \mathcal{P} \mid p \geqslant q\}.$$

Call the subcomplexes $|\mathcal{P}_{\leqslant p}|$ *faces* and the $|\mathcal{P}_{\geqslant q}|$ *cofaces*. So we have two different decompositions of $|\mathcal{P}|$, into faces or into cofaces. In the case at hand, $\Sigma(W, S) = |W\mathcal{S}^f|$, its faces are Coxeter cells and its cofaces are intersections of chambers.

2.5. *Applications and examples*

(a) *Two dimensional complexes.* Let L be a finite graph and m an integer $\geqslant 2$. Let k be the girth of L (the length of the shortest circuit). If $m = 2$, then we assume $k \geqslant 4$. Let $S = Vert(L)$ (the vertex set of L) and define a Coxeter matrix by

$$m_{ss'} = \begin{cases} 1 & \text{if } s = s', \\ m & \text{if } \{s, s'\} \text{ spans an edge}, \\ \infty & \text{otherwise}. \end{cases}$$

Let W be the resulting Coxeter group. Our assumption implies that $N(W, S) = L$. Thus, $\Sigma(W, S)$ is a $CAT(0)$, PE 2-complex such that each 2-cell is a regular $2m$-gon and such that the link of each vertex is L. Here the condition that L was $CAT(1)$ was just that $k(\pi - \pi/m) \geqslant 2\pi$ (which holds provided $k \geqslant 4$ if $m = 2$). We can give Σ a piecewise hyperbolic (abbreviated PH) structure by declaring each 2-cell to be a small regular $2m$-gon in $\mathbb{H}^2$. Since the angles of such a $2m$-gon will be slightly less than in the Euclidean case, we will be able to do this so that links are $CAT(1)$ provided that $k(\pi - \pi/m) > 2\pi$. This holds provided $k > 4$ if $m = 2$ and $k > 3$ if $m = 3$. Thus, provided the condition holds, Σ can be given a PH structure which is $CAT(-1)$.

These "regular" 2-complexes can be thought of as generalization of well known examples of regular tessellations of $\mathbb{E}^2$ and $\mathbb{H}^2$.

Nadia Benakli has made a detailed study of these 2-complexes in her thesis [7]. For example, she shows that the ideal boundary $\Sigma(\infty)$ is usually a Menger curve. Benakli also has another construction of such 2-complexes where the 2-cells are n-gons, with n odd, provided that there is a group G of automorphisms of L such that L'/G is an interval (L' is the barycentric subdivision).

(b) *Word hyperbolic Coxeter groups.* In [51] Moussong also analyzed when the idea of the previous subsection (of replacing the Euclidean Coxeter cells of Σ by hyperbolic Coxeter cells) works in higher dimensions.

Consider the following condition $(*)$ on a Coxeter system (W, S).

$(*)$ For any subset T of S neither of the following holds:
(1) $W_T = W_{T_1} \times W_{T_2}$ with both factors infinite,
(2) W_T is a Euclidean Coxeter group with $\mathrm{Card}(T) \geqslant 3$.

Here a "Euclidean Coxeter group" means the Coxeter group of an orthogonal affine reflection group on $\mathbb{E}^n$ with compact quotient. The "Coxeter diagrams" of these groups are listed in [13, p. 199].

THEOREM 2.22 (Moussong). *The following conditions are equivalent.*
(i) *(W, S) satisfies $(*)$,*
(ii) *Σ can be given a PH, $CAT(-1)$ structure,*
(iii) *W is word hyperbolic,*
(iv) *W does not contain a subgroup isomorphic to $\mathbb{Z} + \mathbb{Z}$.*

To show (i) $\Rightarrow$ (ii) one wants to replace the cells of Σ by Coxeter cells in $\mathbb{H}^n$. In order for the links to remain $CAT(1)$, one needs to know that the length of every closed geodesic in $N(W, S)$ is *strictly* greater than 2π and that the same condition holds for the link of each simplex in $N(W, S)$. In his proof of Lemma 1.28, Moussong analyzed exactly when a metric flag complex has closed geodesics of length equal to 2π. In the case at hand, it was only when conditions (1) and (2) of $(*)$ hold. The implications (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) $\Rightarrow$ (i) are all either well-known or obvious.

(c) *Buildings.* As we mentioned in Section 1.3(d) associated to each building B there is a Coxeter system (W, S) so that each "apartment" is isomorphic to a complex associated to (W, S). Traditionally, this complex is the Coxeter complex (where each chamber is a simplex). For general Coxeter groups, however, it seems more appropriate to use the complex Σ'. Here Σ' denotes the barycentric subdivision of Σ. (In [24] we call this the "modified Coxeter complex".) This point is made by Ronan in [55, p. 184]. The buildings which arise in nature (in algebra or geometry) are usually of spherical or Euclidean type. This means that the Coxeter group W is either finite or Euclidean. In the case where W is Euclidean and irreducible and our definition agrees with the traditional one.

In the general case let us agree that the correct definition of a building should be as a simplicial complex such that each apartment is isomorphic to Σ'. Since, by Theorem 2.8, Σ' can be given the structure of a $CAT(0)$, PE cell complex, we get an induced PE structure on the building B. As explained in Section 1.3(d)), the axioms for buildings imply this structure is $CAT(0)$. Thus, we have the following result, the details of which can be found in [32].

THEOREM 2.23. *Any building (correctly defined) has the structure of a PE simplicial complex which is* $CAT(0)$.

(d) *When is Σ a manifold? A homology manifold?* Since the link of every vertex in Σ is isomorphic to N, these questions can be answered as follows.

PROPOSITION 2.24.
(i) *Σ is a homology n-manifold if and only if N is a homology $(n-1)$-manifold with the homology of S^{n-1}.*
(ii) *For $n \geqslant 5$, Σ is a topological n-manifold if and only if N is as in* (i) *and N is simply connected.*
(iii) *Σ is a PL n-manifold if and only if N is PL homeomorphic to S^{n-1}.*

Statements (i) and (iii) are just restatements of the definitions and statement (ii) follows from Edwards' Theorem 1.52.

We begin by discussing some examples of (iii) when N is a PL triangulation of S^{n-1}.

EXAMPLE 2.25 (*Lanner* [49]). Suppose that $N(W, S)$ is isomorphic to the boundary complex of an n-simplex. Then $n = \text{Card}(S) - 1$, W is infinite, and for every proper subset T of S, the group W_T is finite. Such groups were classified in 1950 by Lanner: they are either irreducible Euclidean reflection groups or hyperbolic reflection groups. In both cases a fundamental chamber is an n-simplex. In the Euclidean case there are a few families in each dimension n and a few exceptional cases in dimensions $\leqslant 8$. In the hyperbolic case, in dimension 2, we have the hyperbolic triangle groups: these are the groups such that $\text{Card}(S) = 3$ and the 3 entries p, q, r of the Coxeter matrix above the diagonal satisfy $(1/p) + (1/q) + (1/r) < 1$. Furthermore, there are 9 hyperbolic examples in dimension 3, 5 more in dimension 4, and none in dimensions > 4. Complete lists can be found on pp. 133 and 199 of [13].

EXAMPLE 2.26 (*Andreev* [4]). Suppose that N is a triangulation of S^2 as a metric flag complex and that condition $(*)$ of subsection (b) holds. Then Andreev proved that W can be realized (uniquely, up to conjugation by an isometry) as a reflection group on $\mathbb{H}^3$. In fact, he shows that there is a convex cell C^3 in $\mathbb{H}^3$, the polar dual of which is N. (Thus, the faces of C^3 corresponding to s and s' make a dihedral angle of $\pi/m_{ss'}$.) W is the group generated by reflections across the faces of C^3.

EXAMPLE 2.27 ([65,31] *and* [21]). Suppose L is the boundary complex of a n-dimensional octahedron (i.e., L is the n-fold join of S^0 with itself). Let W' be a finite Coxeter group of rank n. Use W' to label the edges of one $(n-1)$-simplex in L. Label the other edges 2. This defines a Coxeter group W, with $N(W, S) = L$. Each chamber of W on Σ is combinatorially equivalent to the cone on the dual cellulation of N, i.e., each chamber is a combinatorial cube. If $W' = (\mathbb{Z}/2)^n$, then each chamber actually is an n-cube.

In the case where $W' = S_{n+1}$, the symmetric group, the W-manifolds Σ arise in nature. For example, there is an obvious homomorphism $W \to W' \times (\mathbb{Z}/2)^n$ with kernel Γ_0. Tomei showed in [65], the n-manifold Σ/Γ_0 can be identified with the manifold of all tridiagonal, symmetric $(n+1) \times (n+1)$ matrices with constant spectrum (of distinct eigenvalues). This is also explained in [31]. From a completely direction, it is shown in [21] that for a certain torsion-free, finite-index, normal subgroup Γ_1 of W, Σ/Γ_1 can be identified with Gromov's "Möbius band" hyperbolization construction applied to boundary of a $(n+1)$-cube, and that for a different subgroup Γ_2, Σ/Γ_2 can be identified with the "product with interval" hyperbolization construction applied to the boundary of a simple regular $(n+1)$-cell.

EXAMPLE 2.28. Let L be any flag complex which is a PL triangulation of S^{n-1}, let (W, S) be the corresponding right-angled Coxeter system with $N = L$.

PROPOSITION 2.29. *Hopf's Conjecture* (*from Section* 1.6) *for PE cubical manifolds is equivalent to Conjecture* 1.59.

PROOF. We saw in Section 1.6 that Conjecture 1.56 implies Hopf's Conjecture for PE cubical manifolds. Let L be an arbitrary triangulation of S^{2n-1} by a flag complex and let X^{2n} be the manifold constructed in Corollary 2.18. By the Combinatorial Gauss–Bonnet Theorem (Theorem 1.56),

$$\chi(X^{2n}) = \sum P(L) = [W : \Gamma]P(L).$$

Hence, $\chi(X^{2n})$ and $P(L)$ have the same sign. □

EXAMPLE 2.30. Suppose that R is a ring and that L is a flag complex which is a R-homology $(n-1)$-manifold with the same homology (over R) as S^{n-1} and let (W, S) be the right-angled Coxeter system with $N = L$. For example, we could take $R = \mathbb{Z}[\frac{1}{m}]$ and L to be the lens space $S^{2k-1}/(\mathbb{Z}/m)$ or the suspension of such a lens space. Then Σ is an R-homology n-manifold. Moreover, the ideal boundary $\Sigma(\infty)$ has the same homology (over R) as does S^{n-1} (Compare Theorems 1.50 and 1.52). It follows that W is a virtual Poincaré duality group over R in the sense that any torsion-free subgroup Γ of finite index in W satisfies Poincaré duality over R. However, as is shown in [33], if L is not an integral homology $(n-1)$-sphere, then neither is $\Sigma(\infty)$. Hence, for such an L, W is not a virtual Poincaré duality groups over $\mathbb{Z}$.

(e) *Cohomological dimension.* The *cohomological dimension* of a torsion-free group Γ over a ring R, denoted by $cd_R(\Gamma)$, is nearly the same thing as the smallest dimension of an R-acyclic CW complex on which Γ can act freely. (If the ring R is $\mathbb{Z}$, then we omit it from our notation.) If a group G is not torsion-free, but contains a finite-index, torsion-free subgroup Γ, then the virtual cohomological dimension of G, denoted $vcd_R(G)$, is defined by $vcd_R(G) = cd_R(\Gamma)$. If Γ acts freely and cocompactly on a contractible complex Ω, then $cd_R(\Gamma) = \sup\{n \mid H_c^n(\Omega; R) \neq 0\}$. (See in [16, p. 209].)

EXAMPLE 2.31 (*Bestvina and Mess* [12]). Suppose that L is a flag complex homeomorphic to $\mathbb{R}P^2$, that (W, S) is the right-angled Coxeter system with $N = L$, and that $\Sigma = \Sigma(W, S)$. By Theorem 1.52, $\Sigma(\infty)$ is an inverse limit of connected sums of an increasing number of projective planes. In other words, $\Sigma(\infty)$ is an inverse limit of nonorientable surfaces of increasing genus. Thus, $H^n(\Sigma(\infty); \mathbb{Q}) = 0$ for $n \geqslant 2$, while $H^2(\Sigma(\infty), \mathbb{Z}) = \mathbb{Z}/2$. Consequently, $H_c^n(\Sigma; \mathbb{Q}) = 0$ for $n \geqslant 3$, and $H_c^3(\Sigma; \mathbb{Z}) = \mathbb{Z}/2$. It follows that $vcd_{\mathbb{Q}}(W) = 2$ and $vcd(W) = 3$.

EXAMPLE 2.32 (*Dranishnikov* [38], *Dicks and Leary* [37]). There are similar examples using other Moore spaces than $\mathbb{R}P^2$. For example, suppose a flag complex L_2 is 2-dimensional and has $H_1(L_2; \mathbb{Z}) = \mathbb{Z}/3$ and $\widetilde{H}_i(L_2; \mathbb{Z}) = 0$, for $i \neq 1$. Let (W_2, S_2) be the corresponding Coxeter system. Let (W_1, S_1) be the Coxeter system of the previous example. Put $\Sigma_1 = \Sigma(W_1, S_1)$, $\Sigma_2 = \Sigma(W_2, S_2)$ and $\Sigma = \Sigma_1 \times \Sigma_2$. So, $\dim \Sigma = 6$. Since $\Sigma(\infty)$ is the join of $\Sigma_1(\infty)$ and $\Sigma_2(\infty)$, the Kunneth formula implies that $H^5(\Sigma(\infty); \mathbb{Z}) = \mathbb{Z}/2 \otimes \mathbb{Z}/3 = 0$. Thus, $vcd(W_1) = vcd(W_2) = 3$, while $vcd(W_1 \times W_2) = 5$. Hence, cohomological dimension is not additive for direct products.

Since W acts on the contractible complex Σ with finite isotropy groups, we always have that $vcd(W) \leqslant \dim \Sigma$. As we shall see below, the inequality can be strict.

Let X be a CW complex and $(X_s)_{s \in S}$ a family of subcomplexes as in Section 2.4. For any subset T of S put

$$X_T = \bigcap_{s \in T} X_s$$

and $X_\emptyset = X$. Suppose that the following two conditions hold:

(i) $X_T = \emptyset$, $T \notin \mathcal{S}^f$,

(ii) X_T is acyclic, $T \in \mathcal{S}^f$.

It then follows from Theorem 10.1 in [29] that $\mathcal{U}(W, X)$ is acyclic and, by construction, that W acts on $\mathcal{U}(W, X)$ with finite isotropy groups. Hence, $vcd(W) \leqslant \dim X$. In [9] Bestvina shows that, in fact, $vcd(W)$ is equal to the smallest possible dimension of such an X. In [38] Dranishnikov gives an explicit construction of such a minimal X. (Of course, in most cases, for example if $N(W, S)$ carries some top-dimensional homology, then $vcd(W) = \dim \Sigma$.)

(f) *The Eilenberg–Ganea Problem.* The *geometric dimension* of a torsion-free group Γ, denotes $gd(\Gamma)$, is the smallest dimension of a $K(\Gamma, 1)$ complex. Equivalently, it is the smallest dimension of a contractible CW complex on which Γ can act freely. The following result is proved in [40] when $cd(\Gamma) \neq 1$ and in [60] in the case $cd(\Gamma) = 1$.

THEOREM 2.33 (Eilenberg and Ganea, Stallings).

(i) *If* $cd(\Gamma) \neq 2$, *then* $cd(\Gamma) = gd(\Gamma)$.

(ii) *If* $cd(\Gamma) = 2$, *then either* $gd(\Gamma) = 2$ *or* $gd(\Gamma) = 3$.

The Eilenberg–Ganea Problem is the question of whether or not there exists a group Γ with $cd(\Gamma) = 2$ and $gd(\Gamma) = 3$.

Let L be 2-dimensional flag complex which is (a) acyclic and (b) not simply connected. Let (W, S) be the corresponding right-angled Coxeter system with $N = L$. If X denotes the geometric realization of $\mathcal{S}^f$ then, as in Example 2.21, $\mathcal{U}(W, X) = \Sigma$. Let X_0 be the geometric realization of $\mathcal{S}^f_{>\emptyset}$ (so that X_0 is the barycentric subdivision of L and X is the cone on X) and put $X_{0s} = X_s$ and $\Sigma_0 = \mathcal{U}(W, X_0)$. By the last paragraph of subsection (e), Σ_0 is an acyclic 2-complex. It follows that for any torsion-free subgroup Γ of finite index in W, $cd(\Gamma) = 2$. On the other hand, the only natural contractible complex on which Γ acts is Σ, which has dimension 3. This leads to the following

CONJECTURE 2.34 (*Bestvina*). For any Γ as above, $gd(\Gamma) = 3$.

The reason for believing this is that it seems that X_0 should embed in the universal cover $E\Gamma$ of any $K(\Gamma, 1)$. Furthermore, since $\pi_1(X_0)$ is not trivial it should be impossible to embed it in any contractible 2-complex, since it should not be possible to kill $\pi_1(X_0)$ by adding the same number of 1- and 2-cells. (This is related to the Kervaire Conjecture: if G is a nontrivial group, then any group obtained from G by adding one generator and one relation is also nontrivial.)

(g) *Is Σ homeomorphic to $\mathbb{R}^n$?*

PROPOSITION 2.35.
(i) *If N is a PL triangulation of S^{n-1}, then Σ is PL homeomorphic to $\mathbb{R}^n$.*
(ii) *If N is a PL homology sphere (i.e., N is a PL manifold) and $\pi_1(N)$ is not trivial, then Σ is not simply connected at infinity. (However, Σ is not a manifold, rather it is only a homology manifold.)*
(iii) *If N is a simply connected homology manifold with the homology of S^{n-1}, then, for $n \geqslant 5$, Σ is a contractible manifold.*

PROOF. Statement (i) follows from Stone's result (Theorem 1.47), statement (ii) from Proposition 1.52, and statement (iii) from Edwards' Theorem 1.53. □

We are now in position to tackle Questions 1.45 and 1.46.

PROPOSITION 2.36 ([35, p. 383]). *For each $n \geqslant 5$, there is a Coxeter system (W, S) so that the corresponding PE polyhedron Σ is*
(a) *$CAT(0)$,*
(b) *a topological n-manifold, and*
(c) *not homeomorphic to $\mathbb{R}^n$.*

PROOF. Let A^{n-1} be a compact acyclic PL manifold with boundary such that (1) $\pi_1(\partial A) \to \pi_1(A)$ is onto and (2) the double of A along ∂A is not simply connected. Take a triangulation of A as a flag complex so that ∂A is a full subcomplex. Let N be the simplicial complex resulting from attaching the cone on ∂A to A. Let (W, S) be the right-angled Coxeter system with $N(W, S) = N$. By (1), N is simply connected, so by

Proposition 2.35(iii), Σ is a topological n-manifold. Let (W_1, S_1) be the right-angled Coxeter system whose nerve is the double of A along ∂A. It follows from (2) and part (ii) of Proposition 2.35 that the resulting contractible complex is not simply connected at infinity. On the other hand, W_1 can be identified with an index two subgroup of W. (Double the fundamental chamber X of Σ along X_s where s corresponds to the cone point in N.) Since the fundamental group at infinity of Σ depends only on W (or W_1) we see that Σ is also not simply connected at infinity and hence, not homeomorphic to $\mathbb{R}^n$. □

In [2] it is proved that in the case where $N(W, S)$ is a nonsimply connected PL homology sphere (as in Proposition 2.35(ii)), a modified version of Σ can be taken to be a topological manifold. The idea is similar to that in the previous proof: one blows up the PL singularities of Σ from isolated vertices into intervals. More precisely, we have the following result.

PROPOSITION 2.37 ([2]). *Let L^{n-1} be a PL homology sphere. Then there is a right-angled Coxeter system (W, S), with $N(W, S) = L$, and a PE cubical complex Σ_1 with W-action such that* (a) *Σ_1 is $CAT(0)$ and* (b) *Σ_1 is a topological n-manifold.*

IDEA OF PROOF. We can find a codimension-one homology sphere $L_0 \subset L$ such that (1) L_0 divides L into two pieces L_1 and L_2 (each of which is an acyclic manifold with boundary) and (2) $\pi_1(L_0) \to \pi_1(L_i)$ is onto, for $i = 1, 2$. Triangulate L as a flag complex so that L_0 is a full subcomplex and let (W, S) be the right-angled Coxeter system such that $N(W, S) = L$. For $i = 1, 2$, let N_i denote the union of L_i with the cone over L_0. Then N_i is simply connected. In the construction of Σ a fundamental chamber is essentially a cubical cone over L. As explained in [2], to construct Σ_1, one uses as a chamber the union of two cubical cones: one over N_1 and the other over N_2. These cones are glued together along the cone on L_0. Such a chamber now has PL singularities along an interval which connects the two cone points c_1 and c_2. The link of c_i is N_i, which is simply connected; hence, by Edwards' Theorem 1.53, Σ_1 is a topological manifold. □

REMARK 2.38. In the previous two propositions we have shown that Questions 1.45(a) and 1.46(a) have negative answers. As for part (b), let Σ_1 be as in the previous proposition and consider the $CAT(0)$ manifold $X = \Sigma_1 \times \mathbb{R}$. The group $W \times \mathbb{Z}$ acts on X with compact quotient. Since X is simply connected at infinity, it is homeomorphic to $\mathbb{R}^{n+1}$ (by [59]). On the other hand, $X(\infty)$ is the join of $\Sigma_1(\infty)$ and S^0; hence, it is not homeomorphic to S^n. Question 1.45(c) asks if the ideal boundary is a manifold, then must it be a sphere? It is proved in [3] that the answer is no. The construction is very similar to the one in the above proof of Proposition 2.37. Indeed, if N_1 and N_2 are as in the above proof, then one takes two Euclidean cones (or two hyperbolic cones) on S_1 and N_2 and glue them together along the cone on L_0. The resulting space Y^n is a complete $CAT(0)$ manifold (it is $CAT(-1)$ if we use hyperbolic cones) and $Y(\infty)$ is the original PL homology sphere L^{n-1}. On the other hand, it seems unlikely (because of [68]) that one could construct such an example with a cocompact group of isometries. Thus, the answer to Question 1.46(c) is probably "yes".

3. Artin groups

As explained in Section 3.1 any Coxeter group (finite or not) has a representation as a reflection group on a real vector space. Take the complexification of this vector space. It contains a certain convex open subset such that after deleting the reflection hyperplanes, we obtain an open manifold M on which the Coxeter group W acts freely. The fundamental group of M/W is the "Artin group" A associated to W. When W is finite, Deligne proved that M/W is a $K(A,1)$-space. The conjecture that this should always be the case, is here called the "Main Conjecture". The purpose of this chapter is to outline some work on this conjecture in [24] and [25].

Associated to the Artin group there is a cell complex Φ (which is very similar to Σ). It turns out (Corollary 3.30) that proving the Main Conjecture for W is equivalent to showing Φ is contractible. The complex Φ has a natural PE structure, which we conjecture is always $CAT(0)$. We do not know how to prove this; however, in Section 3.5 we show that there is a (less natural) cubical structure on Φ and that in "most cases" it is $CAT(0)$. Hence, the Main Conjecture holds in most cases.

3.1. *Hyperplane complements*

Let S_n denote the symmetric group on n letters. It acts on $\mathbb{R}^n$ by permutation of coordinates. In fact, this action is as an orthogonal reflection group: the reflections are the transpositions (ij), $1 \leqslant i < j \leqslant n$, the corresponding reflection hyperplanes are the $H_{ij} = \{x \in \mathbb{R}^n \mid x_i = x_j\}$. Complexifying we get an action of S_n on $\mathbb{C}^n = \mathbb{R}^n \otimes \mathbb{C}$ such that S_n acts freely on

$$M = \mathbb{C}^n - \bigcup (H_{ij} \otimes \mathbb{C}).$$

Thus, M/S_n is the configuration space of unordered sets of n distinct points in $\mathbb{C}$.

It is a classical fact that the fundamental group of M/S_n is B_n, the braid group on n strands. The following result is also classical.

THEOREM 3.1 (Fox and Neuwirth [41]).

(1) *M is a $K(\pi,1)$ space, where π is PB_n the pure braid group (i.e., PB_n is the kernel of $B_n \to S_n$).*

(2) *M/S_n is a $K(\pi,1)$ space, for $\pi = B_n$.*

Next suppose that W is a finite reflection group on $\mathbb{R}^n$ and that

$$M = \mathbb{C}^n - \bigcup_r H_r \otimes \mathbb{C},$$

where the union is over all reflections r in W (i.e., all conjugates of elements in S) and where H_r is the hyperplane fixed by r. Arnold and Brieskorn asked if the analogous result to Theorem 3.1 holds in this context. In [36], Deligne proved that this was indeed the case.

THEOREM 3.2 (Deligne [36]). *Suppose that W is a finite reflection group. Then M/W is a $K(\pi,1)$ space, where π is the "Artin group" associated to W (as defined below).*

Artin groups. Suppose that (W, S) is a Coxeter system and that $M = (m_{ss'})$ is the associated Coxeter matrix. Introduce a new set of symbols $X = \{x_s \mid s \in S\}$, one for each element of s.

NOTATION. If m is an integer $\geqslant 2$, then let $\text{prod}(x, y; m)$ denote the word: $xyx\cdots$, where there are a total of m in letters in a word.

DEFINITION 3.3. The *Artin group* associated to (W, S) (or to M) is the group generated by X and with relations:

$$\text{prod}(x_s, x_{s'}; m_{ss'}) = \text{prod}(x_{s'}, x_s; m_{ss'}),$$

where (s, s') range over all elements of $S \times S$ such that $s \neq s'$ and $m_{ss'} \neq \infty$.

REMARK 3.4. If we add the relations $(x_s)^2 = 1$, then the relation appearing in the previous definition can be rewritten as $(x_s x_{s'})^{m_{ss'}} = 1$; hence, we recover the standard presentation of W. Thus, if A is the Artin group associated to (W, S), we see that there is a canonical surjection $p : A \to W$ which send x_s to s.

EXAMPLE 3.5. If W is S_n, then the associated Artin group is B_n.

It is natural to ask if Theorem 3.2 holds in the case where W is infinite. In order to make sense of this question we first need to discuss what is meant by a "linear reflection group" in the infinite case.

Linear reflection groups. Let V be a finite dimensional real vector space. A *linear reflection* on V means a linear involution with fixed space a hyperplane.

Suppose that C is a convex polyhedral cone in V (Figure 5) and that S is a finite set which indexes the set of codimension-one faces of C. Thus, $(C_s)_{s \in S}$ will be the family of codimension-one faces of C. Let H_s denote the linear hyperplane spanned by C_s.

For each $s \in S$, choose a reflection ρ_s with fixed subspace H_s. Let W denote the subgroup of $GL(V)$ generated by $\{\rho_s \mid s \in S\}$.

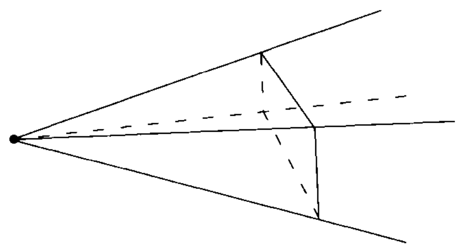


Fig. 5.

DEFINITION 3.6. W is a *linear reflection group* if $w\mathring{C} \cap \mathring{C} = \emptyset$ or all $w \in W$, $w \neq 1$. (Here $\mathring{C}$ denotes the interior of C.)

DEFINITION 3.7. Let

$$\overline{I} = \bigcup_{w \subset W} wC$$

and let I denote the interior of $\overline{I}$. $\overline{I}$ is called the *Tits cone*.

EXAMPLE 3.8. Consider the quadratic form model of $\mathbb{H}^2$: the hyperbolic plane is identified with one sheet of a hyperboloid in $\mathbb{R}^{2,1}$ (3-dimensional Minkowski space). An isometric reflection on $\mathbb{H}^2$ extends to a linear reflection on $\mathbb{R}^{2,1}$ preserving the indefinite quadratic form. Now suppose that W is the reflection group on $\mathbb{H}^2$ generated by the reflections across the edges of a hyperbolic polygon with angles of the form π/m. Then W can be regarded as a linear reflection group on $\mathbb{R}^{2,1}$. (See Figure 6.)

In this case the interior I of the Tits cone is just the interior of the light cone.

THEOREM 3.9 (Vinberg [66]). *Suppose that $W \subset GL(V)$ is a linear reflection group with fundamental polyhedral cone C. Put $C^f = \{x \in C \mid W_x$ is finite$\}$. Then*

(i) *(W, S) is a Coxeter system,*
(ii) *$\overline{I}$ is a convex cone,*
(iii) *I is W-stable and W acts properly on it,*
(iv) *$I \cap C = C^f$,*
(v) *the poset of face of C^f is $\mathcal{S}^f$ (where $\mathcal{S}^f = \{T \subset S \mid W_T$ is finite$\}$).*

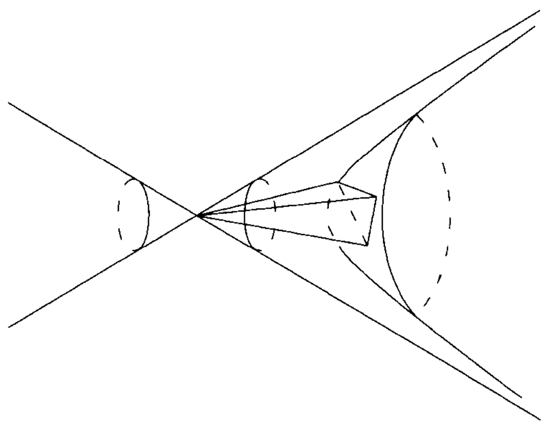


Fig. 6.

Let W be as in Vinberg's Theorem and consider the domain $V + \mathrm{i}I$ in $V \otimes \mathbb{C}$ ($V + \mathrm{i}I$ denotes the set of vectors whose imaginary part lies in I). Set

$$M = (V + \mathrm{i}I) - \bigcup (H_r \otimes \mathbb{C}).$$

The following is the main conjecture which we shall be concerned with in this chapter. According to [50] it is due to Arnold, Pham and Thom.

CONJECTURE 3.10 (*The "Main Conjecture"*). *M/W is a $K(\pi, 1)$ space, where $\pi = A_W$, the Artin group associated to (W, S).*

Some progress on this was made in the thesis of van der Lek [50], where the following result is proved. (Another proof can be found in [24].)

PROPOSITION 3.11 (van der Lek). *$\pi_1(M/W) = A_W$.*

The Main Conjecture can also be formulated in terms of the cell complex Σ which was introduced in the previous chapter in Section 2.3. In fact, in view of Theorem 3.9(v), the following lemma is not surprising.

LEMMA 3.12 ([24, Section 1.2]). *There are W-isovariant homotopy equivalences*:

$$I \sim \Sigma \quad \textit{and} \quad V + \mathrm{i}I \sim I \times I \sim \Sigma \times \Sigma.$$

Set

$$Y = (\Sigma \times \Sigma) - \bigcup_r \Sigma_r \times \Sigma_r,$$

where Σ_r denotes the subcomplex of Σ' fixed by r. Then

$$Y/W \sim M/W.$$

Hence, we have the following

CONJECTURE 3.13 (*Reformulation of the Main Conjecture*). *Y/W is a $K(A_W, 1)$.*

3.2. *The Salvetti complex*

In this section, which is independent of the last three sections of this chapter, we describe a PE cell complex $\widetilde{\Sigma}$ homotopy equivalent to M. The quotient space $\widetilde{\Sigma}/W$ is a finite CW complex. Hence, when the Main Conjecture holds, $\widetilde{\Sigma}/W$ will be a $K(A_W, 1)$ space. The complete details of this construction are given in [25].

In Section 2.1, we considered two posets:

$$\mathcal{S}^f = \{T \subset S \mid W_T \text{ is finite}\} \quad \text{and} \quad W\mathcal{S}^f = \coprod_{T \in \mathcal{S}^f} W/W_T.$$

Here we consider a third poset $W \times \mathcal{S}^f$. The partial ordering on $W \times \mathcal{S}^f$ is defined as follows: $(w, T) < (w', T')$ if and only if

(i) $T < T'$,

(ii) $w^{-1}w' \in W_{T'}$, and

(iii) for all $t \in T$, $\ell(w^{-1}w') < \ell(tw^{-1}w')$ (where ℓ denotes word length in W).

There is a natural projection $\pi : W \times \mathcal{S}^f \to W\mathcal{S}^f$ defined by $(w, T) \to wW_T$. Conditions (i) and (ii) just mean that π is order-preserving. Condition (iii) comes out of the proof of Proposition 3.14, below.

The quickest way to define $\widetilde{\Sigma}$ is to first define its barycentric subdivision $\widetilde{\Sigma}'$; it is the geometric realization of the derived complex of $W \times \mathcal{S}^f$. One then observes that the union of simplices with maximal vertex is (w, T) can be identified with a Coxeter cell of type W_T.

If Z is a cell complex then $\mathcal{P}(Z)$ denotes the poset of cells in Z. For example, $\mathcal{P}(\Sigma) = W\mathcal{S}^f$.

PROPOSITION 3.14 (Salvetti [56], Charney and Davis [25]). *There is a PE cell complex $\widetilde{\Sigma}$ such that*

(i) $\mathcal{P}(\widetilde{\Sigma}) = W \times \mathcal{S}^f$,

(ii) *each cell of $\widetilde{\Sigma}$ is a Coxeter cell,*

(iii) *W acts freely on $\widetilde{\Sigma}$,*

(iv) *$\widetilde{\Sigma}$ is W-equivariantly homotopy equivalent to M (or to Y) and hence, $\widetilde{\Sigma}/W \sim M/W$.*

SKETCH OF PROOF. First, for each (w, T) in $W \times \mathcal{S}^f$ we will describe two open sets in Σ'. Let $U'_{(w,T)}$ denote the open star of the vertex corresponding to wW_T in Σ'. Let $U''_{(w,T)}$ denote the intersection of the open "half spaces" in Σ' which are bounded by the Σ_r with r a reflection in wW_Tw^{-1} and which contain the vertex w. ($U''_{(w,T)}$ is an open "sector".) We note that $U''_{(w,T)}$ contains no point in $U'_{(w,T)}$ with nontrivial isotropy group.

Consider $Y = (\Sigma \times \Sigma) - \bigcup(\Sigma_r \times \Sigma_r)$. Let $U_{(w,T)} = U'_{(w,T)} \times U''_{(w,T)}$. One checks easily that (a) $U_{(w,T)} \subset Y$, (b) $\{U_{(w,T)}\}$ is an open cover of Y, (c) each nonempty intersection of elements in this cover is contractible, and (d) the nerve of this cover is $\widetilde{\Sigma}'$ (the geometric realization of $(W \times \mathcal{S}^f)'$). The proposition follows. □

REMARK. In [56] Salvetti carries out the above construction for arbitrary hyperplane complements. The special case above is done in [25]. When W is the symmetric group, the result was known earlier (for example, to J. Milgram and C. Squier).

The CW complex $\widetilde{\Sigma}/W$ has one cell of dimension Card(T) for each $T \in \mathcal{S}^f$. In particular, when W is finite, $\widetilde{\Sigma}/W$ is the CW complex formed by identifying faces of a single

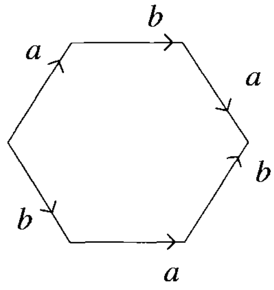


Fig. 7.

Coxeter cell: the precise identifications can be worked out using conditions (i), (ii) and (iii) in the definition of the partial order on $W \times S^f$.

COROLLARY 3.15. *The Main Conjecture holds if and only if $\widetilde{\Sigma}/W$ is a $K(A_W, 1)$ space.*

EXAMPLE 3.16. The Main Conjecture holds when W is finite. In particular, $\widetilde{\Sigma}/W$ is a $K(A, 1)$ when A is an Artin group. For example, when $A = B_3$, a $K(B_3, 1)$ can be constructed by identifying edges of a hexagon in the pattern shown in Figure 7.

EXAMPLE 3.17. If $W = (\mathbb{Z}/2)^n$, then $\widetilde{\Sigma}/W$ can be identified with the n-torus T^n with its usual cell structure (obtained by identifying opposite faces of the n-cube).

EXAMPLE 3.18. Suppose W is right-angled. The Main conjecture holds for W (see Theorem 3.35 below). In this case $\widetilde{\Sigma}/W$ can be identified with a subcomplex of T^n, $n = \text{Card}(S)$, as follows. Let $(z_s)_{s \in S}$ be the standard linear coordinates on T^n. Given a point $z = (z_s)_{s \in S}$ in T^n define the *support of* z by $\text{Supp}(z) = \{s \in S \mid z_s \not\equiv 0 \pmod 1\}$. Then $\widetilde{\Sigma}/W$ can be identified with the subcomplex of T^n consisting of all $z \in T^n$ such that $\text{Supp}(z)$ is the vertex set of a simplex in $N(W, S)$. Thus, $\widetilde{\Sigma}/W$ is a union of subtori, one for each simplex in $N(W, S)$.

In [11], Bestvina and Brady made use of such "right-angled Artin groups" to construct some startling examples of non-finitely presented groups. Suppose that $N(W, S)$ is a finite acyclic complex which is not simply connected (as in 2.5(f)). Let A be the associated Artin group and $\varphi : A \to \mathbb{Z}$ the homomorphism which sends each generator x_s to a generator of $\mathbb{Z}$ and let H denote the kernel of φ. Let $f : \widetilde{\Sigma}/W \to S^1$ be the restriction of the linear map $T^n \to S^1$ which sends z to $\Sigma z_s \pmod 1$. Then f induces φ on π_1. It is proved in [11] that (a) H is not finitely presented and (b) for any $\theta \in S^1$, the inverse image of $f^{-1}(\theta)$ in the universal cover of $\widetilde{\Sigma}/W$ is acyclic. Such H are the first examples of torsion-free groups which are (a) not finitely presented and (b) act freely and cocompactly on acyclic complexes.

COROLLARY 3.19 ([25]). *Suppose the Main Conjecture holds for (W, S). Then*
(i) $cd(A_W) = \dim \widetilde{\Sigma} = \dim \Sigma$.
(ii) $\chi(A_W) = \chi(\widetilde{\Sigma}/W) = 1 - \chi(N(W, S))$.

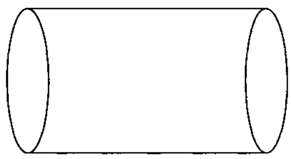


Fig. 8.

A naive idea for proving the Main Conjecture would be to show that $K(\tilde{\Sigma}) \leqslant 0$. This actually works when W is right-angled (as is proved in [25]); moreover, this fact plays a key role in the work of Bestvina and Brady. However, it does not work in general. For example, when $A = B_3$, the link of a vertex in the complex in Example 3.16 is the graph shown in Figure 8.

If each edge length is $2\pi/3$, but then the digons have length $4\pi/3$ which is $< 2\pi$.

3.3. *Complexes of groups*

Graphs of groups. We begin by recalling some well-known results from the theory of graphs of groups (cf. [57]).

Let Ω be a graph, $\mathcal{P}(\Omega)$ the poset of cells in Ω and $\mathcal{P}(\Omega)^{op}$ the dual poset, thought of as a category. A *graph of groups over Ω* is a functor $\mathcal{G}$ from $\mathcal{P}(\Omega)^{op}$ to the category of groups and monomorphisms. Thus, to each vertex v of Ω we are given a group $\mathcal{G}(v)$ and similarly a group $\mathcal{G}(e)$ for each edge e. Moreover, if v is a vertex of e, then there is a monomorphism $\mathcal{G}(e) \to \mathcal{G}(v)$.

From these data one can construct a group G, called the *fundamental group of $\mathcal{G}$* and denoted by $\pi_1(\mathcal{G})$. The basic result in the theory is the following.

THEOREM 3.20 ([57]). *Given a graph of groups $\mathcal{G}$ over Ω, there exists a tree T with G-action ($G = \pi_1(\mathcal{G})$) so that the following hold.*

(i) $T/G = \Omega$.

(ii) *Suppose e is an edge of Ω, v a vertex of e, $\tilde{e}$ a lift of e to T and $\tilde{v}$ the corresponding vertex of $\tilde{e}$. Then there is an isomorphism $G_{\tilde{v}} \cong \mathcal{G}(v)$ taking $G_{\tilde{e}}$ onto the image of $\mathcal{G}(e)$.*

One consequence of (ii) is that the natural map $\mathcal{G}(v) \to \pi_1(\mathcal{G})$ is injective (since it is isomorphic to the inclusion $G_{\tilde{v}} \subset G$). In the language of [46] this means that $\mathcal{G}$ is *developable*. The tree T is called the *universal cover* of $\mathcal{G}$. It is unique up to G-isomorphism. The other feature of a graph of groups is that this universal cover is not only simply connected, it is contractible.

An important application of this theory is to the problem of gluing together various $K(\pi, 1)$ spaces and then being able to decide if the result is also aspherical.

DEFINITION 3.21. An *aspherical realization* of a graph of groups $\mathcal{G}$ is a CW complex B and a map $p: B \to \Omega'$ so that for each vertex v of Ω, $p^{-1}(\text{Star}(v))$ is a $K(\mathcal{G}(v), 1)$. Here "Star" refers to the open star of a vertex in the barycentric subdivision Ω'. (Actually, this is only an approximation of the correct definition which can be found in [47, Sections 3.3 and 3.4].)

REMARK 3.22. It is proved in [47], in the more general context of complexes of groups, that aspherical realizations exist and are unique up to homotopy.

We will usually denote an aspherical realization of $\mathcal{G}$ by $B\mathcal{G}$ and call it the *classifying space* of $\mathcal{G}$. As a definition of $\pi_1(\mathcal{G})$ we could take the usual fundamental group of $B\mathcal{G}$.

The following result is classical; its proof goes back to J.H.C. Whitehead.

THEOREM 3.23. *Let $\mathcal{G}$ be a graph of groups, $B\mathcal{G}$ an aspherical realization, and $G = \pi_1(\mathcal{G})$. Then $B\mathcal{G}$ is a $K(G, 1)$.*

PROOF. Let EG be the universal cover of $K(G, 1)$. Consider the diagonal G-action on $EG \times T$. Projection on the second factor $EG \times T \to T$ induces a map of quotient spaces $EG \times_G T \to \Omega$ which is clearly an aspherical realization. Hence, we can take $B\mathcal{G} = EG \times_G T$. On the other hand, the universal cover of $EG \times_G T$ is $EG \times T$ which is contractible. □

Complexes of groups. Here we present a simplified version of the theory developed in [46] and [47]. (For the applications we have in mind we do not need the most general version of the theory.)

DEFINITION 3.24. Let $\mathcal{P}$ be a poset. A *simple complex of groups over $\mathcal{P}$* is a functor $\mathcal{G}$ from $\mathcal{P}$ to the category of groups and monomorphisms.

REMARK. In the general situation of [47], $\mathcal{P}$ need not be a poset but only a "category without loop". More importantly, $\mathcal{G}$ need not be a functor. The appropriate triangular diagrams relating compositions of morphisms need not commute on the nose but only up to conjugation by some elements in the target group; furthermore, these elements must be kept track of.

The concept of an "aspherical realization" is defined as before. Such an aspherical realization is devoted by $B\mathcal{G}$ and called the *classifying space of $\mathcal{G}$*. By definition, $\pi_1(\mathcal{G}) = \pi_1(B\mathcal{G})$. It can also be defined via generators and relations [47, §12.8]. A complex of groups need not be developable. Moreover, even if it is developable its universal cover need not be contractible.

EXAMPLE 3.25. Suppose $\mathcal{P} = \mathcal{P}(\Delta^2)^{op}$ where Δ^2 is a 2-simplex. Define a complex of groups $\mathcal{G}$ by Figure 9, where D_m denotes the dihedral groups of order $2m$. Then $\pi_1(\mathcal{G})$ is a Coxeter group W on three generators. $\mathcal{G}$ is developable. Assume that $(1/p) + (1/q) + (1/r) > 1$. Then W is finite. The universal cover of $\mathcal{G}$ is homeomorphic to S^2 (triangulated

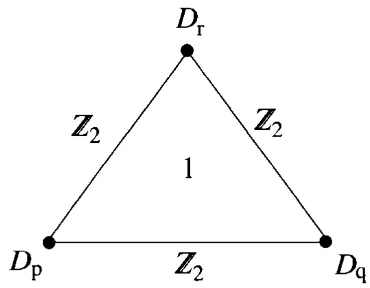


Fig. 9.

as the Coxeter complex) and $B\mathcal{G} = EW \times_W S^2$ which is *not* a $K(W, 1)$ (its universal cover is homotopy equivalent to S^2).

What is missing in higher dimensions is a hypothesis of nonpositive curvature (which is automatic in the case of a graph of groups). The proof of the following basic result is outlined by Haefliger in [46] and some of the details are supplied in the thesis of Spieler [58].

THEOREM 3.26 (Haefliger, Spieler). *Let $\mathcal{G}$ be a complex of groups. Suppose that $\mathcal{G}$ admits "a metric with $K \leqslant 0$". Then*
(i) *$\mathcal{G}$ is developable,*
(ii) *its universal cover is $CAT(0)$ (and hence, contractible),*
(iii) *$B\mathcal{G}$ is a $K(G, 1)$ space.*

REMARK. Suppose $\mathcal{G}$ is a complex of groups over a poset $\mathcal{P}$. The hypothesis of nonpositive curvature means that the geometric realization of $\mathcal{P}'$ admits a PE "orbihedral structure" with $K \leqslant 0$. In other words, the local models of a universal cover must be PE and $CAT(0)$.

EXAMPLE 3.27. Suppose (W, S) is a Coxeter system and that $\mathcal{S}^f$ is the poset defined in Section 2.1. Define a simple complex of groups $\mathcal{W}$ over $\mathcal{S}^f$ to be the functor $\mathcal{W}(T) = W_T$. Then it is easily seen that $\mathcal{W}$ is developable and that $\pi_1(\mathcal{W}) = W$. Moreover, its universal cover is just the geometric realization of the poset $W\mathcal{S}^f$, i.e., it is Σ' (the barycentric subdivision of Σ). Moussong's Theorem 2.8 shows that the natural PE structure on Σ' is $CAT(0)$, i.e., we are in the situation of Theorem 3.26.

EXAMPLE 3.28. Let A be the Artin group associated to (W, S). Let $\mathcal{S}^f$ be as above. Define a simple complex of groups $\mathcal{A}$ over $\mathcal{S}^f$ to be the functor $\mathcal{A}(T) = A_T$, where A_T is the Artin group corresponding to (W_T, T) (A_T is an Artin group of "finite type"). It follows easily that $\pi_1(\mathcal{A}) = A$. Moreover, it can be shown that for each $T \in \mathcal{S}^f$, $A_T \to A$ is injective, i.e., $\mathcal{A}$ is developable.

Consider the set,

$$A\mathcal{S}^f = \coprod_{T \in \mathcal{S}^f} A/A_T,$$

partially ordered by inclusion. The geometric realization Φ of $\mathcal{AS}^f$ will be called the *modified Deligne complex.* It will be our principal object of interest in the remaining two sections. As is in the case of Σ', it is easy to see that Φ is simply connected and therefore, that it is the universal cover of $\mathcal{A}$.

3.4. *Reinterpretation of the Main Conjecture*

Recall that

$$Y = (\Sigma \times \Sigma) - \bigcup \Sigma_r \times \Sigma_r.$$

Let X denote the geometric realization of $\mathcal{S}^f$. Let $\pi : Y/W \to \Sigma'/W = X$ be the map induced by projection on the first factor.

PROPOSITION 3.29 ([25]). *$\pi : Y/W \to X$ is an aspherical realization of $\mathcal{A}$ (from Example 3.28).*

PROOF. Basically, this is just what Deligne's Theorem (Theorem 3.2) tells us. Indeed, if T is a vertex of X, then

$$\pi^{-1}\big(\mathrm{Star}(T)\big) = \Big[\mathrm{Star}(1, T) \times \Sigma - \bigcup \mathrm{hyperplanes}\Big] / W_T$$

which is homotopy equivalent the orbit space of the hyperplane complement for the finite Coxeter group W_T. By Deligne's result, this is a $K(A_T, 1)$. Thus, in the general situation, Y/W is homotopy equivalent to $B\mathcal{A}$. □

COROLLARY 3.30. *The Main Conjecture holds for (W, S) if and only if Φ is contractible.*

PROOF. We are using the form of the Main Conjecture in 3.13. By Proposition 3.29, Y/W is homotopy equivalent to $EA \times_A \Phi$. Therefore, the universal cover of Y/W is contractible if and only if Φ is contractible. □

Φ is "building-like". As explained in Section 3.1, there is a natural epimorphism $p : A \to W$. We can define a section $\varphi : W \to A$ of p as follows. Given w in W, write $w = s_1 \cdots s_m$ where $s_1 \cdots s_m$ is a word of minimum length for w. Set $\varphi(w) = x_{s_1} \cdots x_{s_m} \in A$. It can be shown that φ is well-defined (i.e., the value of $\varphi(w)$ does not depend on the choice of minimal word). Of course, φ is not a homomorphism. The map φ induces a embedding of posets $W\mathcal{S}^f \to \mathcal{AS}^f$ and therefore, a simplicial embedding $\Sigma' \to \Phi$. The translates of Σ' by elements of A are the *apartment-like* subcomplexes. We have that

$$\Phi = \bigcup_{a \in A} a\Sigma'$$

and in this sense, Φ is "building-like". Φ is not in fact a building: two points need not lie in a common apartment.

Nevertheless, there is an obvious idea for trying to prove Φ is contractible. Give Φ a PE structure by declaring each apartment-like complex to be isometric to Σ' with its natural PE structure (described in Sections 2.2 and 2.3). Then prove Φ is $CAT(0)$. To attack this we must study links of vertices in Φ.

The simplicial complexes $\widehat{\Sigma}$ and $\widehat{\Phi}$. Suppose for the moment that W is a finite Coxeter group. Let Δ be a simplex, the codimension-one faces of which are indexed by S: if $s \in S$, then Δ_s denotes the corresponding face. Given $x \in \Delta$, put $S(x) = \{s \in S \mid x \in \Delta_s\}$. Define

$$\widehat{\Sigma}_W = (W \times \Delta)/\sim,$$

where the equivalence relation $\sim$ is defined by $(w, x) \sim (w', x')$ if and only if $x = x'$ and $w^{-1}w' \in W_{S(x)}$. $\widehat{\Sigma}_W$ is the usual *Coxeter complex* of W. Its poset of simplices is

$$\left(\coprod_{T \neq S} W/W_T\right)^{op}.$$

It can be identified with the triangulation of the unit sphere in $\mathbb{R}^n$ where the $(n-1)$ simplices are the intersection of S^{n-1} with the translates of a fundamental simplicial cone. In other words, if we identify Δ with a spherical simplex so that the dihedral angle along $\Delta_s \cap \Delta_{s'}$ is $\pi/m_{ss'}$, then the induced PS structure on $\widehat{\Sigma}_W$ is that of a round sphere.

Similarly, define

$$\widehat{\Phi}_W = (A \times \Delta)/\sim,$$

where $\sim$ is defined by $(a, x) \sim (a', x')$ if and only if $x = s'$ and $a^{-1}a' \in A_{S(x)}$. The simplicial complex $\widehat{\Phi}_W$ is called the *Deligne complex* of (W, S). Deligne proved, in [36], that $\widehat{\Phi}_W$ is homotopy equivalent to a wedge of $(n-1)$-spheres (where $n - 1 = \dim \Delta$).

As in the previous subsection,

$$\widehat{\Phi}_W = \bigcup_{a \in A} a\widehat{\Sigma}_W.$$

So, $\widehat{\Phi}_W$ is spherical-building-like. If we identify Δ with a spherical simplex as above, then each apartment-like subcomplex $a\widehat{\Sigma}_W$ is isometric to a round sphere.

Links of vertices in Φ. The vertices of Φ correspond to elements of $A\mathcal{S}^f$, i.e., to cosets of the form aA_T, $T \in \mathcal{S}^f$. We classify these into three types.

Type 1. $T = \emptyset$. In this case $Lk(v, \Phi) = N(W, S)$ (the same link as for a vertex of Σ).

Type 2. T is maximal in $\mathcal{S}^f$. In this case, $Lk(v, \Phi) = \widehat{\Phi}_{W_T}$. (In the analogous case for Σ' the link would be a round sphere.)

Type 3. In the general case, $Lk(v, \Phi)$ is the orthogonal join of a link of a simplex in $N(W, S)$ and a link of Type 2.

Thus, if we give Φ its natural PE structure every link is of the form $N(W, S)$ (which is $CAT(1)$ by Moussong's Lemma and Corollary 2.16), or a Deligne complex associated to a finite Coxeter group, or a join of these two types. Thus, Φ will be $CAT(0)$ provided the following holds.

CONJECTURE 3.31. Let W be a finite Coxeter group. Then the Deligne complex $\widehat{\Phi}_W$ with its round metric is $CAT(1)$.

THEOREM 3.32. *Conjecture* 3.31 *implies the Main Conjecture* 3.10.

It follows from a lemma of [5, p. 210] that Conjecture 3.31 holds when $W = D_m$, the dihedral group of order $2m$. (The lemma asserts that $\widehat{\Phi}_{D_m}$ has no circuits of length $\leqslant 2m$.) This yields the following.

COROLLARY 3.33 ([25]). *The Main Conjecture holds whenever* $\dim \Phi = 2$.

3.5. *A cubical structure on* Φ

A Coxeter cell of type W_T can be subdivided into combinatorial cubes, so that cubes containing the barycenter of the cell correspond to the simplices of $\widehat{\Sigma}_{W_T}$. (See Figure 10.) Given an arbitrary Coxeter system (W, S), we could give $\Sigma'(W, S)$ a cubical structure by declaring each such combinatorial cube to be regular Euclidean cube. As in the previous section, there are three types of vertices to consider. For those of Type 2 (where T is maximal), $Lk(v, \Sigma')$ is $\widehat{\Sigma}_{W_T}$ with its all right PS structure. By [17, p. 29], the Coxeter complex of a finite Coxeter group is a flag complex; hence, by Gromov's Lemma, the all right structure on $\widehat{\Sigma}_{W_T}$ is $CAT(1)$. Consider a vertex of Type 1 (where $T = \emptyset$). The link is $N(W, S)$ with an all right PS structure. Hence, this link is $CAT(1)$ if and only if $N(W, S)$ is a flag complex. Since the links of Type 3 are orthogonal joins of versions of Type 1 and 2, we see that the cubical structure on Σ is $CAT(0)$ if and only if $N(W, S)$ is a flag complex.

In exactly the same way, we can put a cubical structure on Φ. In the case of vertices of Type 2, we have the following key lemma of [24, Lemma 4.3.2].

LEMMA 3.34. *Let W be a finite Coxeter group. Then the Deligne complex $\widehat{\Phi}_W$ is a flag complex.*

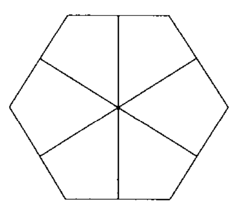


Fig. 10.

For vertices of Type 1 the link is again $N(W, S)$. The conclusion is that the cubical structure on Φ is $CAT(0)$ exactly when the cubical structure on Σ is $CAT(0)$.

THEOREM 3.35 ([25]). *The Main Conjecture is true when $N(W, S)$ is a flag complex.*

REMARK 3.36. Taken together, Corollary 3.33 and Theorem 3.35 constitute a proof of the Main Conjecture in most cases.

REMARK 3.37. In [49], Lanner showed that $N(W_T, T)$ is the boundary of a simplex (an "empty simplex"in $N(W, S)$) if and only if W_T is a reflection group on either hyperbolic space or Euclidean space with fundamental chamber a simplex. (See Example 2.25.) Hence, $N(W, S)$ is a flag complex if and only if for all subsets T of S, with $\text{Card}(T) \geqslant 3$, neither of the following conditions hold:
(a) W_T is an irreducible, affine Euclidean reflection group,
(b) W_T is a hyperbolic reflection group with fundamental chamber a simplex.

References

[1] A.D. Aleksandrov, V.N. Berestovskii and I. Nikolaev, *Generalized Riemannian spaces*, Russian Math. Surveys **41** (1986), 1–54.
[2] F. Ancel, M. Davis and C. Guilbault, *CAT(0) reflection manifolds*, Geometric Topology (Athens, GA, 1993), AMS/IP Stud. Adv. Math. 2.1, Amer. Math. Soc., Providence, RI (1997), pp. 441–445.
[3] F. Ancel and C. Guilbault, *Interiors of compact contractible N-manifolds are hyperbolic* ($N \geqslant 5$), J. Differential Geom. **45** (1997), 1–32.
[4] E.M. Andreev, *On convex polyhedra in Lobacevskii spaces*, Math. USSR Sbornik **10** (1970), 413–440.
[5] K.I. Appel and P.E. Schupp, *Artin groups and infinite Coxeter groups*, Invent. Math. **72** (1983), 201–220.
[6] W. Ballmann, M. Gromov and V. Schroeder, *Manifolds of Nonpositive Curvature*, Progress in Math., Vol. 61, Birkhäuser, Stuttgart (1985).
[7] N. Benakli, *Polyedres hyperboliques passage du local au global*, Thèse, Université de Paris-Sud (1992).
[8] V.N. Berestovskii, *Borsuk's problem on the metrization of a polyhedron*, Soviet Math. Dokl. **27** (1983), 56–59.
[9] M. Bestvina, *The virtual cohomological dimension of Coxeter groups*, Geometric Group Theory I, G. Niblo and M. Roller, eds, Cambridge University Press, Cambridge (1993).
[10] M. Bestvina, *The local homology properties of boundaries of groups*, Michigan Math. J. **43** (1996), 123–129.
[11] M. Bestvina and N. Brady, *Morse theory and finiteness properties of groups*, Invent. Math. **129** (1997), 123–129.
[12] M. Bestvina and G. Mess, *The boundary of negatively curved groups*, J. Amer. Math. Soc. **4** (1991), 469–481.
[13] N. Bourbaki, *Groupes et Algebrès de Lie*, Chapters IV–VI, Masson, Paris (1981).
[14] M.R. Bridson, *Geodesics and curature in metric simplicial complexes*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific, Singapore (1991), 373–463.
[15] M. Bridson and A. Haefliger, *Metric Spaces of Nonpositive Curvature*, Springer, New York (1999).
[16] K. Brown, *Cohomology of Groups*, Springer, New York (1982).
[17] K. Brown, *Buildings*, Springer, New York (1989).
[18] M. Brown, *Some applications of an approximation theorem for inverse limits*, Proc. Amer. Math. Soc. **11** (1960), 478–481.
[19] J.W. Cannon, *Shrinking cell-like decompositions of manifolds: codimension three*, Ann. of Math. **110** (1979), 83–112.

[20] R. Charney and M.W. Davis, *Singular metrics of nonpositive curvature on branched covers of Riemannian manifolds*, Amer. J. Math. **115** (1993), 929–1009.

[21] R. Charney and M.W. Davis, *Strict hyperbolization*, Topology **34** (1995), 329–350.

[22] R. Charney and M.W. Davis, *The Euler characteristic of a nonpositively curved, piecewise Euclidean manifold*, Pacific J. Math. **171** (1995), 117–137.

[23] R. Charney and M.W. Davis, *The polar dual of a convex polyhedral set in hyperbolic space*, Michigan Math. J. **42** (1995), 479–510.

[24] R. Charney and M.W. Davis, *The $K(\pi, 1)$-problem for hyperplane complements associated to infinite reflection groups*, J. Amer. Math. Soc. **8** (1995), 597–627.

[25] R. Charney and M.W. Davis, *Finite $K(\pi, 1)$'s for Artin groups*, Prospects in Topology, F. Quinn, ed., Annals of Math. Studies, Vol. 138, Princeton Univ. Press, Princeton (1995), 110–124.

[26] R. Charney, M. Davis and G. Moussong, *Nonpositively curved, piecewise Euclidean structures on hyperbolic manifolds*, Michigan Math. J. **44** (1997), 201–208.

[27] J. Cheeger, W. Müller and R. Schrader, *On the curvature of piecewise flat spaces*, Comm. Math. Phys. **92** (1984), 405–454.

[28] S.S. Chern, *On curvature and characteristic classes of a Riemannian manifold*, Abh. Math. Sem. Univ. Hamburg **20** (1956), 117–126.

[29] M.W. Davis, *Groups generated by reflections and aspherical manifolds not covered by Euclidean space*, Ann. of Math. (2) **117** (1983), 293–325.

[30] M.W. Davis, *Coxeter groups and aspherical manifolds*, Algebraic Topology Aarhus 1982, Lecture Notes in Math., Vol. 1051, Springer, New York (1984), 197–221.

[31] M.W. Davis, *Some aspherical manifolds*, Duke Math. J. **55** (1987), 105–139.

[32] M.W. Davis, *Buildings are $CAT(0)$*, Geometry and Cohomology in Group Theory, P. Kropholler, G. Niblo and R. Stohr, eds, London Math. Soc. Lecture Notes, Vol. 252, Cambridge Univ. Press, Cambridge (1998), 108–123.

[33] M.W. Davis, *The cohomology of a Coxeter group with group ring coefficients*, Duke Math. J. **91** (1998), 297–314.

[34] M.W. Davis and J.-C. Hausmann, *Aspherical manifolds without smooth or PL structure*, Algebraic Topology, Lecture Notes in Math., Vol. 1370, Springer, New York (1989).

[35] M.W. Davis and T. Januszkiewicz, *Hyperbolization of polyhedra*, J. Differential Geom. **34** (1991), 347–388.

[36] P. Deligne, *Les immeubles de groupes de tresses généralisés*, Invent. Math. **17** (1972), 273–302.

[37] W. Dicks and I.J. Leary, *On subgroups of Coxeter groups*, Geometry and Cohomology in Group Theory, P. Kropholler, G. Niblo and R. Stohr, eds, London Math. Soc. Lecture Notes, Vol. 252, Cambridge Univ. Press, Cambridge (1998), 124–160.

[38] A.N. Dranishnikov, *The virtual cohomological dimension of Coxeter groups*, Proc. Amer. Math. Soc. **125** (1997), 1885–1891.

[39] R.D. Edwards, *The topology of manifolds and cell-like maps*, Proc. ICM Helsinki (1978), 111–127.

[40] S. Eilenberg and T. Ganea, *On the Lusternik–Schnirelman category of abstract groups*, Ann. of Math. **65** (1957), 517–518.

[41] R.H. Fox and L. Neuwirth, *The braid groups*, Math. Scand. **10** (1962), 119–126.

[42] M. Freedman, *The topology of four-dimensional manifolds*, J. Differential Geom. **17** (1982), 357–453.

[43] R. Geroch, *Positive sectional curvature does not imply positive Gauss–Bonnet integrand*, Proc. Amer. Math. Soc. **54** (1976), 267–270.

[44] E. Ghys and P. de la Harpe, eds, *Sur les groupes hyperboliques d'apres Mikhael Gromov*, Prog. Math., Vol. 83, Birkhäuser, Boston (1990).

[45] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, M.S.R.I. Publ., Vol. 8, S.M. Gersten, ed., Springer, New York (1987), 75–264.

[46] A. Haefliger, *Complexes of groups and oribhedra*, Group Theory from a Geometrical Viewpoint, E. Ghys, A. Haefliger and A. Verjousky, eds, World Scientific, Singapore (1991), 504–540.

[47] A. Haefliger, *Extensions of complexes of groups*, Ann. Inst. Fourier, Grenoble **42** (1–2) (1992), 275–311.

[48] M.A. Kervaire, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. **144** (1969), 67–72.

[49] F. Lanner, *On complexes with transitive groups of automorphisms*, Comm. Sem. Math. Univ. Lund **11** (1950), 1–71.

[50] H. van der Lek, *The homotopy type of complex hyperplane complements*, Ph.D. Thesis, University of Nijmegan (1983).
[51] G. Moussong, *Hyperbolic Coxeter groups*, Ph.D. Thesis, The Ohio State University (1988).
[52] C. Plaut, *Metric spaces of curvature* $\geqslant k$, Handbook of Geometric Topology, R.J. Davermann and R.B. Sher, eds, Elsevier, Amsterdam (2002), 819–898.
[53] I. Rivin and C. Hodgson, *A characterization of compact convex polyhedra in hyperbolic* 3*-space*, Invent. Math. **111** (1993), 77–111.
[54] D. Rolfsen, *Strongly convex metrics in cells*, Bull. Amer. Math. Soc. **78** (1968), 171–175.
[55] M. Ronan, *Lectures on Buildings*, Academic Press, San Diego (1989).
[56] M. Salvetti, *Topology of the complement of real hyperplanes in* $\mathbb{C}^n$, Invent. Math. **88** (1988), 603–618.
[57] J.P. Serre, *Trees*, Springer, New York (1989).
[58] B. Spieler, *Nonpositively curved orbihedra*, Ph.D. Thesis, The Ohio State University (1992).
[59] J.R. Stallings, *The piecewise-linear structure of Euclidean space*, Proc. Cambridge Phil. Soc. **58** (1962), 481–488.
[60] J.R. Stallings, *On torsion-free groups with infinitely many ends*, Ann. of Math. **88** (1968), 312–334.
[61] R. Stanley, *Flag* f*-vectors and the cd-index*, Math. Z. **216** (1994), 483–499.
[62] D. Stone, *Geodesics in piecewise linear manifolds*, Trans. Amer. Math. Soc. **215** (1976), 1–44.
[63] W. Thurston, *The geometry and topology of* 3*-manifolds*, Princeton University (1979) (reproduced lecture notes).
[64] P. Thurston, *The topology of* 4*-dimensional G-spaces and a study of* 4*-manifolds of non-positive curvature*, Ph.D. Thesis, University of Tennessee (1993).
[65] C. Tomei, *The topology of the isospectral manifold of tridiagonal matrices*, Duke Math. J. **51** (1984), 981–996.
[66] E.B. Vinberg, *Discrete linear groups generated by reflections*, Math. USSR Izvestija **5** (5) (1971), 1083–1119.
[67] D.W. Walkup, *The lower bound conjecture for* 3*- and* 4*-manifolds*, Acta Math. **125** (1970), 75–107.
[68] D. Wright, *Contractible open manifolds which are not covering spaces*, Topology **31** (1992), 281–291.

CHAPTER 9

Cohomological Dimension Theory

Jerzy Dydak*

University of Tennessee, Knoxville, TN 37996, USA

E-mail: dydak@math.utk.edu

Contents

*Supported in part by grant DMS-9704372 from the National Science Foundation.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Abstract

The paper outlines the development of the cohomological dimension theory from its beginning (around 1930) to the present. The emphasis is on the main ideas and on the interaction between geometric intuition and the power of algebra. The historical development of ideas is mentioned with an explanation from the modern point of view.

0. Historical development of ideas

0.1. *Dimension*

One of the most intuitive geometric concepts is that of dimension. While it is very easy for a non-mathematician to guess the dimension of a particular geometric object, it is fairly difficult to define the dimension in a formal, rigorous way. The first mathematical definition of the dimension arises in linear algebra and one easily gets that the dimensions of the line, the plane, and the n-space (consisting of ordered n-tuples of real numbers) are respectively 1, 2, and n. After the formation of topology as an offshoot of analysis, mathematicians attempted to define dimension in a non-algebraic way. The most intuitive is the so-called *small inductive dimension* $\operatorname{ind}(X)$ of a space X. Essentially, $\operatorname{ind}(X) \leqslant n$ means that open sets U with $\operatorname{ind}(\operatorname{Fr} U) \leqslant n-1$ form a basis of X.

0.1. DEFINITION OF SMALL INDUCTIVE DIMENSION. Given a regular space X, one assigns to X the *small inductive dimension* $\operatorname{ind} X$ (also called the Menger–Urysohn dimension) which is an integer larger than or equal to -1, or infinity (∞), by the following conditions:

(1) $\operatorname{ind}(X) = -1$ if and only if X is empty,
(2) $\operatorname{ind}(X) \leqslant n$, where $0 \leqslant n < \infty$, if for every point $x \in X$ and each neighborhood $V \subset X$ of x there is an open set $U \subset V$ in X containing x so that $\operatorname{ind}(\operatorname{Fr} U) \leqslant n-1$,
(3) $\operatorname{ind}(X) = n$, where $0 \leqslant n < \infty$, if $\operatorname{ind}(X) \leqslant n$ and $\operatorname{ind}(X) \leqslant n-1$ does not hold,
(4) $\operatorname{ind}(X) = \infty$, if $\operatorname{ind}(X) \leqslant n$ does not hold for all $n < \infty$.

Small inductive dimension ind works best for separable metric spaces. For normal spaces one has a traditional definition of the covering dimension dim.

0.2. DEFINITION OF COVERING DIMENSION. Given a normal space X, one assigns to X the *covering dimension* $\dim(X)$ (also called the Čech–Lebesgue dimension) which is an integer larger than or equal to -1, or infinity (∞), by the following conditions:

(1) $\dim(X) \leqslant n$, where $0 \leqslant n < \infty$, if every finite open cover of X has a finite open refinement of order at most n (this means that every $n+2$ distinct elements of the cover have empty intersection),
(2) $\dim(X) = n$, where $0 \leqslant n < \infty$, if $\dim(X) \leqslant n$ and $\dim(X) \leqslant n-1$ does not hold,
(3) $\dim(X) = \infty$, if $\dim(X) \leqslant n$ does not hold for all $n < \infty$.

There is a third way of defining dimension (the *large inductive dimension*) and we refer to [63] for a discussion of cases in which various topological dimensions coincide (it is so for the most significant class of topological spaces: the metrizable and separable spaces). We conclude this section by stating that in the case of Euclidean spaces the topological and algebraic dimensions are equal to each other.

0.3. BROUWER'S THEOREM. *The algebraic and topological dimensions coincide for* R^n.

Historically, Theorem 0.3, also known as the Fundamental Theorem of Dimension Theory, is perhaps the first result relating algebraic/analytic and topological concepts. We refer the reader to [8] for a discussion of the ultimate theorem relating algebraic/analytic and topological ideas: the Atiyah–Singer Index Formula.

0.2. *Cohomological dimension*

With the introduction of homology/cohomology groups it made sense to develop the concept of dimension in an algebraic fashion. The fundamental theorem is due to Alexandroff

0.4. ALEXANDROFF'S THEOREM. *If X is a finite-dimensional compactum, then its dimension is the smallest integer n such that $\check{H}^{n+1}(X, A; \mathbf{Z}) = 0$ for all closed subsets A of X.*

One can replace $\mathbf{Z}$ by any abelian group $G \neq 0$ which leads to the notion of the cohomological dimension $\dim_G(X)$.

0.5. DEFINITION OF COHOMOLOGICAL DIMENSION. Given a paracompact space X and an abelian group G, one assigns to X the *cohomological dimension* $\dim_G(X)$ which is an integer larger than or equal to -1, or infinity (∞), by the following conditions:

(1) $\dim_G(X) \leqslant n$, where $0 \leqslant n < \infty$, if $\check{H}^{n+1}(X, A; G) = 0$ for all closed subsets A of X,
(2) $\dim_G(X) = n$, where $0 \leqslant n < \infty$, if $\dim_G(X) \leqslant n$ and $\dim_G(X) \leqslant n - 1$ does not hold,
(3) $\dim_G(X) = \infty$, if $\dim_G(X) \leqslant n$ does not hold for all $n < \infty$.

The relevance of dimension with respect to arbitrary groups was underscored by Pontryagin [92] who realized that in order to construct compacta X and Y with $\dim(X \times Y) < \dim(X) + \dim(Y)$, one has to construct a compactum P such that $\dim(P) > \dim_{\mathbf{Z}/p}(P)$ for some prime p.

0.6. PONTRYAGIN'S THEOREM. *For each prime $\mathbf{p}$ there is a compactum $P_{\mathbf{p}}$ such that the following conditions hold*:

(1) $\dim(P_{\mathbf{p}}) = 2$, $\dim_{\mathbf{Z}/\mathbf{q}}(P_{\mathbf{p}}) = 1$ *if $\mathbf{q}$ is a prime and $\mathbf{q} \neq \mathbf{p}$*,
(2) $\dim(P_{\mathbf{p}} \times P_{\mathbf{q}}) = 3$ *if $\mathbf{q}$ is a prime and $\mathbf{p} \neq \mathbf{q}$*.

The year 1935 was crucial for the future development of the cohomological dimension theory. During the First International Topology Conference in Moscow, Alexandroff posed several questions (see [2]), and the following two turned out to be of significant importance.

0.7. ALEXANDROFF'S PROBLEM. Is there a countable family $\mathcal{G}$ of abelian groups such that for any compactum X and any abelian group G, the dimension $\dim_G(X)$ can be expressed in terms of $\dim_H(X)$, $H \in \mathcal{G}$?

0.8. ALEXANDROFF'S PROBLEM. Is there an infinite-dimensional compactum X such that $\dim_G(X) < \infty$ for

(a) $G = \mathbf{Z}$,
(b) some $G \neq 0$?

Problem 0.7 was solved by Bockstein [7] in 1956, who singled out the following groups (known now as *Bockstein groups*):

0.9. DEFINITION OF BOCKSTEIN GROUPS. The set $\mathcal{B}_{\mathcal{G}}$ of *Bockstein groups* is

$$\{\mathbf{Q}\} \cup \bigcup_{\mathbf{p}\ \text{prime}} \{\mathbf{Z}/\mathbf{p}, \mathbf{Z}/\mathbf{p}^{\infty}, \mathbf{Z}_{(\mathbf{p})}\},$$

where $\mathbf{Z}/\mathbf{p}^{\infty}$ is the $\mathbf{p}$-torsion of $\mathbf{Q}/\mathbf{Z}$, and $\mathbf{Z}_{(\mathbf{p})}$ consists of the rationals whose denominator is not divisible by $\mathbf{p}$.

In the process of solving Problem 0.7, Bockstein discovered his exact sequence which led to cohomology operations, an important concept in algebraic topology (see [103]). Pontryagin made a serious attempt to solve Problem 0.8. He knew that spheres are closely related to the concept of dimension (see 0.12) and felt that first one needs to get a better understanding of the homotopy groups of spheres. This led to another sizable part of algebraic topology. Thus, cohomological dimension theory is, historically, part of the bedrock of modern algebraic topology.

The first part of the development of the cohomological dimension theory culminates with the publishing of the survey article by Kuzminov [80], which is a 'must read' even today. The second part of the development of the cohomological dimension theory was jump started by the extremely elegant paper of Walsh [107]. Generally speaking, [80] relies on algebra as the driving force, while in [107] it is the geometry which plays the dominant role. In the view of this author, the following aspects of [107] were decisive in awarding it the role of a milestone:

(a) it established a bridge between geometric topology and the cohomological dimension theory in the form of cell-like maps,
(b) it pursued, for the first time, a formal geometric analogy between covering dimension and the integral cohomological dimension,
(c) it introduced a geometric construction, known as the Edwards–Walsh complex, which turned out to be a very useful tool.

The Edwards–Walsh complex provided the geometry in Dranishnikov's [28] solution of Problem 0.8. The algebra consisted of computing the K-theory groups of the Eilenberg–MacLane complex $K(\mathbf{Z}, 3)$ (details later on). By analyzing the cohomology exact sequence of a pair (X, A)

$$\cdots \to \check{H}^n(X; G) \to \check{H}^n(A; G) \to \check{H}^{n+1}(X, A; G) \to \cdots$$

it was known that $\dim_G(X) \leqslant n$ is equivalent to $K(G, n)$ being an absolute extensor of X (see 0.11). First, let us recall the definition of absolute extensors.

0.10. DEFINITION OF ABSOLUTE EXTENSOR. K is an *absolute extensor* of X (notations: $K \in AE(X)$ or $X \tau K$) if every map $f : A \to K$ extends over X if A is closed in X.

0.11. COHEN'S THEOREM. *If X is locally compact, then* $\dim_G(X) \leqslant n$ *if and only if* $K(G, n) \in AE(X)$.

In the case of the covering dimension one has a similar result.

0.12. ALEXANDROFF'S THEOREM. *If X is compact, then* $\dim(X) \leqslant n$ *if and only if* $S^n \in AE(X)$.

The strength of [107] is that statements 0.11–0.12 were chosen as definitions rather than theorems, which resulted in elegance and simplicity. This idea was pursued by Dranishnikov [32], Dydak [47] and Dranishnikov and Dydak [38]. As a result, we are currently at the third stage of the development of the cohomological dimension theory, known as Extension Theory.

We refer the reader to the classic work [68] for an account of the dimension theory of separable metric spaces. The starting point of that book is the small inductive dimension. It contains a proof of 0.3 on p. 41 (for the original proof see [16]) and a proof of 0.12 on p. 83. For the most up to date account of dimension theory, including the relations between various concepts of dimension, we refer to [63].

Since Alexandroff is the founder of cohomological dimension theory, it is not surprising that most early research in that area was done in the former Soviet Union. We recommend papers by Bockstein, Boltyanskij, Kuzminov, Pontryagin, Shvedov, Sitnikov, Sklyarenko, and Zarelua (see our list of references) if one wants to get a taste of the achievements of the Soviet School of Topology. Outside of the former Soviet Union, there was also research done by Borsuk, Cohen, Dyer, Kodama, and Nagata.

As of now, there is no book which covers a sizable part of cohomological dimension theory. There is a set of unpublished notes [40], and [85] contains a chapter devoted to cohomological dimension theory (one can find a proof of 0.4 there). Bredon's book [14] contains some results for cohomological dimension with coefficients in principal ideal domains. We recommend [80] for a discussion of the pre-Dranishnikov era (one can find a proof of 0.11 there), and we recommend [27] for an account of Dranishnikov's solutions of all the major problems of classical cohomological dimension theory. Chigogidze [18] has extended several results of cohomological dimension theory of metrizable spaces to Tychonoff spaces.

1. Bockstein inequalities and Bockstein theorems

To tackle Problem 0.7, Bockstein needed a way to compare the cohomological dimensions of a space with respect to various groups. This was achieved via his exact sequence.

1.1. BOCKSTEIN'S EXACT SEQUENCE. *Suppose X is a topological space and A is its closed subspace. If $0 \to G_1 \to G_2 \to G_3 \to 0$ is a short exact sequence of abelian groups, then there is an exact sequence*

$$\cdots \to H^n(X,A;G_1) \to H^n(X,A;G_2) \\ \to H^n(X,A;G_3) \to H^{n+1}(X,A;G_1) \to \cdots.$$

Bockstein's exact sequence immediately implies a set of inequalities.

1.2. General Bockstein's Inequalities. *Suppose X is a topological space and $0 \to G_1 \to G_2 \to G_3 \to 0$ is a short exact sequence of abelian groups. Then the following inequalities hold*:

(a) $\dim_{G_2}(X) \leqslant \max(\dim_{G_1}(X), \dim_{G_3}(X))$,

(b) $\dim_{G_1}(X) \leqslant \max(\dim_{G_2}(X), \dim_{G_3}(X) + 1)$,

(c) $\dim_{G_3}(X) \leqslant \max(\dim_{G_1}(X) - 1, \dim_{G_2}(X))$.

In the case of Bockstein groups, general Bockstein inequalities can be reduced as follows:

1.3. Bockstein's Inequalities. *Suppose X is a topological space. Then the following inequalities hold*:

(1) $\dim_{\mathbf{Z}/\mathbf{p}^\infty}(X) \leqslant \dim_{\mathbf{Z}/\mathbf{p}}(X)$,

(2) $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant \dim_{\mathbf{Z}/\mathbf{p}^\infty}(X) + 1$,

(3) $\dim_{\mathbf{Q}}(X) \leqslant \dim_{\mathbf{Z}_{(\mathbf{p})}}(X)$,

(4) $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant \dim_{\mathbf{Z}_{(\mathbf{p})}}(X)$,

(5) $\dim_{\mathbf{Z}/\mathbf{p}^\infty}(X) \leqslant \max(\dim_{\mathbf{Q}}(X), \dim_{\mathbf{Z}_{(\mathbf{p})}}(X) - 1)$,

(6) $\dim_{\mathbf{Z}_{(\mathbf{p})}}(X) \leqslant \max(\dim_{\mathbf{Q}}(X), \dim_{\mathbf{Z}/\mathbf{p}^\infty}(X) + 1)$.

Using his inequalities Bockstein was able to solve Problem 0.7.

1.4. First Bockstein's Theorem. *If X is compact, then*

$$\dim_G(X) = \max\{\dim_H(X) \mid H \in \sigma(G)\}.$$

The set $\sigma(G)$ appearing in the theorem above is known as the *Bockstein basis* of G, and is defined as follows:

1.5. Definition of the Bockstein basis. Given an abelian group G its *Bockstein basis* $\sigma(G)$ is the subset of $\mathcal{B}_{\mathcal{G}}$ defined as follows:

(1) $\mathbf{Q} \in \sigma(G)$ if and only if $\mathbf{Q} \otimes G \neq 0$,

(2) $\mathbf{Z}/\mathbf{p} \in \sigma(G)$ if and only if $(\mathbf{Z}/\mathbf{p}) \otimes G \neq 0$,

(3) $\mathbf{Z}_{(\mathbf{p})} \in \sigma(G)$ if and only if $(\mathbf{Z}/\mathbf{p}^\infty) \otimes G \neq 0$,

(4) $\mathbf{Z}/\mathbf{p}^\infty \in \sigma(G)$ if and only if $(\mathbf{Z}/\mathbf{p}^\infty) * G \neq 0$ (here $H * G$ is the torsion product of groups H and G).

Since Pontryagin's example (see Theorem 0.6) demonstrated that the logarithmic law $\dim(X \times Y) = \dim(X) + \dim(Y)$, known to be true in the case of Euclidean spaces, does not hold for compacta, it was natural to seek formulae describing $\dim_G(X \times Y)$ in terms of the cohomological dimensions of X and Y. This problem was completely solved by Bockstein as follows:

1.6. Second Bockstein Theorem. *Suppose X and Y are compact. Then,*

(1) $\dim_{\mathbf{Z}/\mathbf{p}}(X \times Y) = \dim_{\mathbf{Z}/\mathbf{p}}(X) + \dim_{\mathbf{Z}/\mathbf{p}}(Y)$,

(2) $\dim_{\mathbf{Q}}(X \times Y) = \dim_{\mathbf{Q}}(X) + \dim_{\mathbf{Q}}(Y)$,

(3) $\dim_{\mathbf{Z}/\mathbf{p}^\infty}(X \times Y) = \max\{\dim_{\mathbf{Z}/\mathbf{p}^\infty}(X) + \dim_{\mathbf{Z}/\mathbf{p}^\infty}(Y), \dim_{\mathbf{Z}/\mathbf{p}}(X) + \dim_{\mathbf{Z}/\mathbf{p}}(Y) - 1\}$,
(4) $\dim_{\mathbf{Z}_{(\mathbf{p})}}(X \times Y) = \dim_{\mathbf{Z}_{(\mathbf{p})}}(X) + \dim_{\mathbf{Z}_{(\mathbf{p})}}(Y)$ *if* $\dim_{\mathbf{Z}_{(\mathbf{p})}}(X) = \dim_{\mathbf{Z}/\mathbf{p}^\infty}(X)$ *or* $\dim_{\mathbf{Z}_{(\mathbf{p})}}(Y) = \dim_{\mathbf{Z}/\mathbf{p}^\infty}(Y)$, *and*
(5) $\dim_{\mathbf{Z}_{(\mathbf{p})}}(X \times Y) = \max\{\dim_{\mathbf{Z}/\mathbf{p}^\infty}(X \times Y) + 1, \dim_{\mathbf{Q}}(X) + \dim_{\mathbf{Q}}(Y)\}$ *if* $\dim_{\mathbf{Z}_{(\mathbf{p})}}(X) > \dim_{\mathbf{Z}/\mathbf{p}^\infty}(X)$ *and* $\dim_{\mathbf{Z}_{(\mathbf{p})}}(Y) > \dim_{\mathbf{Z}/\mathbf{p}^\infty}(Y)$.

Here is the geometric reason why the cohomological dimension works better in the case of Cartesian products of compacta: $K \in AE(X \times Y)$ is a stronger condition than $K^Y \in AE(X)$ (here, K^Y is the space of maps from Y to K in the compact-open topology). Indeed, any map $f : A \to K^Y$, A being closed in X, induces a map $f' : A \times Y \to K$ via $f'(a, y) = f(a)(y)$. If f' extends over $X \times Y$, then f extends over X. If K is an Eilenberg–MacLane space $K(G, n)$, then K^Y is known to be the product of Eilenberg–MacLane spaces $K(\check{H}^{n-i}(Y; G), i)$, $0 \leqslant i \leqslant n$ (this is a geometric version of the Kunneth Formula, for a more general result see [26]). On the other hand, if K is a sphere (recall that the covering dimension is related to maps into spheres), the space K^Y is much more complicated.

The First Bockstein Theorem means that the cohomological dimension of a compactum X is determined by the set $\{\dim_H(X)\}_{H \in \mathcal{B}_\mathcal{G}}$, and the Bockstein Inequalities impose certain restrictions on this set. This leads to the notion of a Bockstein function.

1.7. DEFINITION OF BOCKSTEIN FUNCTION. A *Bockstein function* is a function $D : \mathcal{B}_\mathcal{G} \to \mathbf{Z}_{\geqslant 0}$ from the set of Bockstein groups to the set of non-negative integers plus infinity such that the following holds:
(1) $D(\mathbf{Z}/\mathbf{p}^\infty) \leqslant D(\mathbf{Z}/\mathbf{p})$,
(2) $D(\mathbf{Z}/\mathbf{p}) \leqslant D(\mathbf{Z}/\mathbf{p}^\infty) + 1$,
(3) $D(\mathbf{Q}) \leqslant D(\mathbf{Z}_{(\mathbf{p})})$,
(4) $D(\mathbf{Z}/\mathbf{p}) \leqslant D(\mathbf{Z}_{(\mathbf{p})})$,
(5) $D(\mathbf{Z}/\mathbf{p}^\infty) \leqslant \max(D(\mathbf{Q}), D(\mathbf{Z}_{(\mathbf{p})}) - 1)$,
(6) $D(\mathbf{Z}_{(\mathbf{p})}) \leqslant \max(D(\mathbf{Q}), D(\mathbf{Z}/\mathbf{p}^\infty) + 1)$,
(7) if $\min D = 0$, then $D \equiv 0$.

Obviously, the basic example of a Bockstein function is the one induced by a compactum.

1.8. DEFINITION OF DIMENSION FUNCTION OF A COMPACTUM. Given a compactum X, its *dimension function* D_X is the Bockstein function defined by $D_X(H) = \dim_H(X)$.

A natural question is:

1.9. BOCKSTEIN–BOLTYANSKI PROBLEM. Is every Bockstein function equal to the dimension function of a compactum?

This problem was completely solved by Dranishnikov (details later on). Notice that, in view of the First Bockstein Theorem 1.4 and the Alexandroff Theorem 0.4, the dimension of a compactum corresponds to the well-known concept of the supremum of a function:

1.10. DEFINITION OF THE DIMENSION OF A BOCKSTEIN FUNCTION. The *dimension* $\dim D$ of a Bockstein function is $\sup\{D(H) \mid H \in \mathcal{B}_{\mathcal{G}}\}$.

It turns out one can translate operations on compact spaces into operations on Bockstein functions.

1.11. DEFINITION OF A WEDGE OF BOCKSTEIN FUNCTIONS. Given a set $\{D_a\}_{a\in A}$ of Bockstein functions, one defines its *wedge* $\bigvee_{a\in A} D_a$ by

$$\bigvee_{a\in A} D_a(H) = \max\{D_a(H) \mid a \in A\}.$$

1.12. DEFINITION OF THE PRODUCT OF BOCKSTEIN FUNCTIONS. Given two Bockstein functions D_1 and D_2, one defines their *product* $D_1 \times D_2$ as follows:

(1) $(D_1 \times D_2)(\mathbf{Z/p}) = D_1(\mathbf{Z/p}) + D_2(\mathbf{Z/p})$,
(2) $(D_1 \times D_2)(\mathbf{Q}) = D_1(\mathbf{Q}) + D_2(\mathbf{Q})$,
(3) $(D_1 \times D_2)(\mathbf{Z/p^\infty}) = \max\{D_1(\mathbf{Z/p^\infty}) + D_2(\mathbf{Z/p^\infty}), D_1(\mathbf{Z/p}) + D_2(\mathbf{Z/p}) - 1\}$,
(4) $(D_1 \times D_2)(\mathbf{Z_{(p)}}) = D_1(\mathbf{Z_{(p)}}) + D_2(\mathbf{Z_{(p)}})$ if $D_1(\mathbf{Z_{(p)}}) = D_1(\mathbf{Z/p^\infty})$ or $D_2(\mathbf{Z_{(p)}}) = D_2(\mathbf{Z/p^\infty})$, and
(5) $(D_1 \times D_2)(\mathbf{Z_{(p)}}) = \max\{(D_1 \times D_2)(\mathbf{Z/p^\infty}) + 1, D_1(\mathbf{Q}) + D_2(\mathbf{Q})\}$ if $D_1(\mathbf{Z_{(p)}}) > D_1(\mathbf{Z/p^\infty})$ and $D_2(\mathbf{Z_{(p)}}) > D_2(\mathbf{Z/p^\infty})$.

In this way one gets an algebra:

1.13. DEFINITION OF BOCKSTEIN ALGEBRA. The *Bockstein algebra* $(\mathcal{BF}, \vee, \times)$ is the set of all Bockstein functions $\mathcal{BF}$ equipped with two associative binary operations $\vee$ and $\times$ so that $\vee$ is distributive with respect to $\times$.

Now, the Second Bockstein Theorem 1.6 can be reformulated as:

1.14. THEOREM. *If X and Y are compact, then*
(1) $D_{X\vee Y} = D_X \vee D_Y$,
(2) $D_{X\times Y} = D_X \times D_Y$.

A more sophisticated reformulation is:

1.15. THEOREM. *There is an epimorphism $X \to D_X$ of the algebra of pointed compact spaces with operations of wedge and smash product to the Bockstein algebra.*

The significance of Theorem 1.15 is that, quite often, one does calculations in the Bockstein Algebra and interprets them in terms of geometrical properties of X.

The best source for all the results of this section is [80]. Notice that 1.5 is taken from [40] and is equivalent to the definition of the Bockstein basis which can be found in [80].

2. Cell-like maps

The connection between cohomological dimension and geometric topology in the form of cell-like mappings was established in [107].

2.1. DEFINITION OF CELL-LIKE SETS/MAPS. A space A is *cell-like* if any map from A to a CW complex is null-homotopic. The map $f: X \to Y$ is *cell-like* if f is proper and $f^{-1}(y)$ is cell-like for all $y \in Y$.

Cell-like sets and maps are more suitable in point set topology, while in geometric topology one has the related concept of *cellular subsets/maps*.

2.2. DEFINITION OF CELLULAR SUBSETS/MAPS. A compact subset A of an n-manifold M^n is *cellular* if open n-cells containing A form a basis of neighborhoods of A. The map $f: M^n \to Y$ is *cellular* if f is proper and $f^{-1}(y)$ is a cellular subset of M^n for all $y \in Y$.

Cellular maps arise in almost every question involving the existence of homeomorphisms between manifolds. The first significant application of the concept of cellularity was demonstrated by M. Brown in his proof of the Generalized Schoenflies Conjecture. The most basic result is that cellular maps between closed manifolds are near-homeomorphisms.

2.3. THEOREM. *Suppose M^n and N^n are closed n-manifolds. The closure of the set of homeomorphisms from M^n to N^n in the compact-open topology coincides with the set of all cellular maps from M^n to N^n.*

In practice, instead of a cell-like map between two manifolds, one has a cell-like decomposition G of a manifold M^n and one wonders if the natural projection $\pi: M^n \to M^n/G$ is a near-homeomorphism. The fundamental result in this area is:

2.4. CELL-LIKE APPROXIMATION THEOREM. *Suppose G is a cell-like decomposition of an n-manifold M^n, where $n \geqslant 5$. Then $\pi: M^n \to M^n/G$ is a near-homeomorphism if and only if M^n/G is finite-dimensional and has the disjoint disks property.*

Thus, one arrives at the following problem.

2.5. CELL-LIKE MAPPING PROBLEM. If $f: X \to Y$ is cell-like and X is finite-dimensional compactum, is Y finite dimensional?

This is the point of overlap of geometric topology and cohomological dimension theory. Namely, [107] contains a characterization of cell-like images of n-dimensional compacta.

2.6. EDWARDS–WALSH THEOREM. *The compactum Y is a cell-like image of an n-dimensional compactum if and only if the integral cohomological dimension of Y is at most n.*

In this way one gets the equivalence of the Cell-like Mapping Problem 2.5 and the Alexandroff Problem 0.8(a).

In point set topology one has a generic construction of cell-like maps (see [69]). We will present this construction in the more general setting of shape theory.

2.7. DEFINITION OF A SHAPE EQUIVALENCE. A map $f : X \to Y$ is called a *shape equivalence* if, for any CW complex K, the induced function $f^* : [Y, K] \to [X, K]$ of sets of homotopy classes is a bijection. f is called a *hereditary shape equivalence* if, for any closed subset A of Y, the map $f|f^{-1}(A) : f^{-1}(A) \to A$ is a shape equivalence.

Notice that X is cell-like if and only if the unique map from X to any one-point space is a shape equivalence. In particular, each hereditary shape equivalence is a cell-like map. We are ready to present the general construction of cell-like maps.

2.8. THEOREM. *For every compactum X there is a hereditary shape equivalence $\pi : \widetilde{X} \to X$ such that $\widetilde{X}$ is the inverse limit of finite simplicial complexes with bonding maps being simplicial.*

PROOF. To a shape theorist, the idea of the construction is quite simple: any compactum is shape equivalent to an inverse sequence $\{\pi_n^{n+1} : X_{n+1} \to X_n\}$ of finite simplicial complexes with bonding maps being simplicial (see [55] or [82]). The inverse limit $\widetilde{X}$ of $\{\pi_n^{n+1} : X_{n+1} \to X_n\}$ is also shape equivalent to $\widetilde{X}$ (see [55] or [82]). Therefore, one would expect X and $\widetilde{X}$ to be shape equivalent via a map. Let us show this in a more detailed manner.

For each $n \geqslant 1$ choose a finite open cover $\mathcal{U}_n$ of X so that the diameter of each element of $\mathcal{U}_n$ is less than 2^{-n} and $\mathcal{U}_{n+1}$ refines $\mathcal{U}_n$. Let X_n be the nerve of $\mathcal{U}_n$ and let $\pi_n^{n+1} : X_{n+1} \to X_n$ be a simplicial map determined by choosing, for each $U \in \mathcal{U}_{n+1}$, an element $U' \in \mathcal{U}_n$ containing U. Let $\widetilde{X}$ be the inverse limit of $\{\pi_n^{n+1} : X_{n+1} \to X_n\}$. Given a point z in X_n, z can be expressed as a linear combination $\sum_{i=1}^k c_i \cdot U_i$, where $U_i \in \mathcal{U}_n$ for $i \leqslant k$, $c_i > 0$ for $i \leqslant k$, $\sum_{i=1}^k c_i = 1$, and $\bigcap_{i=1}^k U_i \neq \emptyset$. We define the carrier $carr(z)$ of z to be the closure of $\bigcap_{i=1}^k U_i$. Notice that the diameter $\operatorname{diam}(carr(z))$ of $carr(z)$ is less than 2^{-n} and $carr(z) \subset carr(\pi_{n-1}^n(z))$. Thus, if $x = \{z_n\}_{n\geqslant 1} \in \widetilde{X}$, then $\bigcap_{n=1}^\infty carr(z_n)$ consists of precisely one point, denoted by $\pi(x)$. It is easy to check that $\pi : \widetilde{X} \to X$ is continuous, so it suffices to show that π is a shape equivalence. (Indeed, $\pi|\pi^{-1}(A) : \pi^{-1}(A) \to A$ can be obtained from A in the same manner as π was obtained from X. Thus, $\pi|\pi^{-1}(A)$ is a shape equivalence for each closed subset A of X, which means that π is a hereditary shape equivalence.)

For each n we have the natural projection $\pi_n : \widetilde{X} \to X_n$, and there is a map $p_n : X \to X_n$ induced by a partition of unity associated with $\mathcal{U}_n$. This means that if $p_n(x) = \sum_{i=1}^k c_i \cdot U_i$, where $U_i \in \mathcal{U}_n$ for $i \leqslant k$, $c_i > 0$ for $i \leqslant k$, and $\sum_{i=1}^k c_i = 1$, then $x \in \bigcap_{i=1}^k U_i$. Both $\{\pi_n\}_{n\geqslant 1}$ and $\{p_n\}_{n\geqslant 1}$ are shape equivalences (see [55] or [82] on how to interpret this statement). Thus, to show that π is a shape equivalence, it suffices to prove that $p_n \circ \pi$ is homotopic to π_n. Suppose $z = \{z_j\}_{j\geqslant 1} \in \widetilde{X}$ and $z_n = \sum_{i=1}^k c_i \cdot U_i$, where $U_i \in \mathcal{U}_n$ for $i \leqslant k$, $c_i > 0$ for $i \leqslant k$, and $\sum_{i=1}^k c_i = 1$. We know that $\pi(z) \in \operatorname{cl}(\bigcap_{i=1}^k U_i)$. Suppose $p_n(\pi(z)) =$

$\sum_{s=1}^{r} d_s \cdot V_s$, where $d_s > 0$ for each s. In view of the way p_n was described, we know that $\pi(z) \in \bigcap_{s=1}^{r} V_s$. Consequently, $(\bigcap_{i=1}^{k} U_i) \cap (\bigcap_{s=1}^{r} V_s) \neq \emptyset$, and $V_1, \ldots, V_r, U_1, \ldots, U_k$ form a simplex of X_n. Now, the formula $H(z,t) = t \cdot \pi(z) + (1-t) \cdot p_n(\pi(z))$ gives a homotopy joining $p_n \circ \pi$ and π. □

Notice that if one defines $\widetilde{X}^{(k)}$ to be the inverse limit of the maps in the k-skeleta, $\pi_n^{n+1} : X_{n+1}^{(k)} \to X_n^{(k)}$, then $\widetilde{X}^{(0)}$ is a Cantor set (unless X is discrete), and $\pi : \widetilde{X}^{(0)} \to X$ is onto. Thus, one recovers the well-known fact that every compactum is an image of a Cantor set. More generally, $\pi : \widetilde{X}^{(k)} \to X$ is a UV^{k-1}-map (see [74] for a definition of UV^k-maps). Thus, every compactum is an image of a k-dimensional compactum via a UV^{k-1}-map.

We refer the reader to [83] for the history of the Cell-like Mapping Problem and Theorem 2.3. A proof of Theorem 2.4 can be found in [21]. See [78,79], and [108] for cases in which the Cell-like Mapping Problem has a positive solution. The Edwards–Walsh Theorem 2.6 has been generalized to non-compact metric spaces by Rubin and Schapiro [96]. Other generalizations involve cohomological dimension with groups different from integers; e.g., Dranishnikov [27] considered integers modulo **p** and compact metric spaces, where the property of a map being cell-like is replaced by the property of being acyclic. The case of non-compact metric spaces was considered by Koyama and Yokoi [76]. A different generalization of the Edwards–Walsh Theorem appears in [74] (the compact metric case) and [75] (the non-compact metric case). Namely, the authors characterize spaces which are images of UV^{n-1} – and acyclic maps (with respect to a ring with unity) of n-dimensional spaces.

3. Edwards–Walsh complexes

As we mentioned (see 0.12), the covering dimension of compacta can be characterized in terms of extension of maps into spheres. There is a dual characterization in terms of approximate lifting of maps. It is well-known that X is at most n-dimensional if and only if any map from X to a CW complex L can be approximated by a map to the n-skeleton $L^{(n)}$ of L. This property of n-dimensional compacta can be restated in terms of approximate liftings. First, we need to introduce the concept of a combinatorial map.

3.1. DEFINITION OF COMBINATORIAL MAPS. A map $\pi : \widetilde{L} \to L$ of CW complexes is called *combinatorial* provided $\pi^{-1}(K)$ is a subcomplex of $\widetilde{L}$ whenever K is a subcomplex of L.

3.2. DEFINITION OF APPROXIMATE LIFTS. Suppose $\pi : \widetilde{L} \to L$ is a combinatorial map and L is a simplicial complex with the CW topology. A map $f : X \to L$ is said to *approximately lift* with respect to π provided there is a map $\tilde{f} : X \to \widetilde{L}$ such that, for each $x \in X$ and each simplex Δ in L, $f(x) \in \Delta$ implies $\pi \circ \tilde{f}(x) \in \Delta$.

Now, we can restate the property of approximating maps by maps into the n-skeleton in terms of approximate lifts.

3.3. THEOREM. *X is at most n-dimensional if and only if, for any simplicial complex L, every map $f: X \to L$ has an approximate lift with respect to the inclusion $L^{(n)} \hookrightarrow L$.*

One of the goals of [107] was to generalize this result to the case of the integral cohomological dimension. This result is mentioned in the abstract of [60] but a proof was never published.

3.4. THEOREM. *Let n be an integer. For each simplicial complex L there is a CW complex $EW_{\mathbf{Z}}(L, n)$ and a combinatorial map $\pi_L : EW_{\mathbf{Z}}(L, n) \to L$ so that the following conditions are equivalent:*

(a) $\dim_{\mathbf{Z}}(X) \leqslant n$,
(b) *for any simplicial complex L, every map $f: X \to L$ has an approximate lift with respect to π_L.*

Before outlining the details of the construction of the Edwards–Walsh complexes $EW_{\mathbf{Z}}(L, n)$, let us discuss a simpler construction which stems from the work of Pontryagin and which can be viewed as a precursor of the Edwards–Walsh construction. To simplify our exposition we will deal with ordered simplicial complexes, i.e., complexes with all their vertices being ordered (for example, by enumeration). Ordered simplicial complexes form a category, with morphisms being simplicial inclusions which preserve the order of vertices. We will consider its subcategory $\mathcal{SC}_2$ consisting of ordered simplicial complexes of dimension at most 2.

3.5. DEFINITION. $\mathcal{SC}$ ($\mathcal{SC}_2$) is the category of ordered simplicial complexes (of dimension at most 2). Morphisms of $\mathcal{SC}$ ($\mathcal{SC}_2$) are simplicial inclusions which preserve the order of vertices. $\mathcal{CW}$ ($\mathcal{CW}_2$) is the category of of CW complexes (of dimension at most 2) with all cells being attached via PL-maps (in this way one can give a PL-structure to any of their finite subcomplexes). Morphisms of $\mathcal{CW}$ ($\mathcal{CW}_2$) are continuous maps which induce an isomorphism of the domain with the image of the map.

3.6. THEOREM. *There is a functor $P_{\mathbf{Z}/\mathbf{p}}$ from $\mathcal{SC}_2$ to $\mathcal{CW}_2$ and there is a natural transformation $\pi_L : P_{\mathbf{Z}/\mathbf{p}}(L) \to L$ of functors with the following properties:*

(a) *Suppose A is a closed subset of an at most 2-dimensional compactum X so that* $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant 1$. *Given an ordered simplicial complex L with* $\dim(L) \leqslant 2$, *given a map $g: A \to P_{\mathbf{Z}/\mathbf{p}}(L)$, and given an extension $f: X \to L$ of $\pi_L \circ g$, there is an extension $G: X \to P_{\mathbf{Z}/\mathbf{p}}(L)$ of g so that G is an approximate lift of f with respect to π_L.*
(b) *Suppose X is an at most 2-dimensional compactum. If any map $f: X \to L$ from X to a 2-dimensional ordered simplicial complex L has an approximate lift with respect to π_L, then* $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant 1$.
(c) *Given an object L of $\mathcal{SC}_2$, given a subcomplex K of L, and given a map $f: K \to K(\mathbf{Z}/\mathbf{p}, 1)$, the map $f \circ \pi_K$ extends over $P_{\mathbf{Z}/\mathbf{p}}(L)$.*

PROOF. If L is of dimension at most 1, then we put $P_{\mathbf{Z}/\mathbf{p}}(L) = L$ and $\pi_L : P_{\mathbf{Z}/\mathbf{p}}(L) \to L$ is the identity map. Now, it suffices to define $P_{\mathbf{Z}/\mathbf{p}}(\Delta)$ for the standard 2-simplex Δ

with vertices v_0, v_1, v_2 so that $v_0 < v_1 < v_2$. We attach a 2-cell Δ' via a PL-map $\partial\Delta' \to \partial\Delta$ of degree p. Thus, Δ' is the 2-skeleton of a $K(\mathbf{Z}/\mathbf{p}, 1)$, with $\partial\Delta$ being its 1-skeleton. Since Δ is contractible, $\partial\Delta' \to \partial\Delta$ extends to $\pi_L : \Delta' \to \Delta$, giving us the map $\pi_L : P_{\mathbf{Z}/\mathbf{p}}(\Delta) = \Delta' \to \Delta$. Suppose L is an ordered simplicial complex of dimension at most 2. Given a 2-simplex s of L there is a unique isomorphism $i_s : \Delta \to s$. $P_{\mathbf{Z}/\mathbf{p}}(L)$ is the union of adjunction spaces $\Delta' \cup_{i_s|\partial\Delta} L^{(1)}$ and $\pi_L : P_{\mathbf{Z}/\mathbf{p}}(L) \to L$ is the unique map so that $\pi_L|\Delta' \cup_{i_s|\partial\Delta} L^{(1)}$ is the identity on $L^{(1)}$ and equals $i_s \circ \pi_\Delta$ on $\Delta' \subset \Delta' \cup_{i_s|\partial\Delta} L^{(1)}$.

Notice that $P_{\mathbf{Z}/\mathbf{p}}(\Delta)$, being the 2-skeleton of of a $K(\mathbf{Z}/\mathbf{p}, 1)$, is an absolute extensor of any, at most 2-dimensional, compactum X with $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant 1$. Indeed, given a closed subset A of X, and given $f : A \to P_{\mathbf{Z}/\mathbf{p}}(\Delta)$, f can be extended to $F : X \to K(\mathbf{Z}/\mathbf{p}, 1)$. Since $\dim(X) \leqslant 2$, F can be approximated, rel. A, by $F' : X \to P_{\mathbf{Z}/\mathbf{p}}(\Delta) = K(\mathbf{Z}/\mathbf{p}, 1)^{(2)}$.

Now, condition (a), in the case $L = \Delta$, follows from the fact that $P_{\mathbf{Z}/\mathbf{p}}(\Delta)$ is an absolute extensor of X. Since any $P_{\mathbf{Z}/\mathbf{p}}(L)$ is built by putting copies of $P_{\mathbf{Z}/\mathbf{p}}(\Delta)$ together, condition (a) follows.

Implication (b) is weaker than implication (c). Before showing it, let us make sure implication (c) holds. As before, it suffices to show it in the case of L being the standard 2-simplex Δ. Suppose K is a subcomplex of Δ and $f : K \to K(\mathbf{Z}/\mathbf{p}, 1)$ is a map. Notice that f can be extended over the 1-skeleton of Δ. Thus, without loss of generality, it suffices to consider the case $K = \partial\Delta$. Since the attaching map $\alpha : \partial\Delta' \to \partial\Delta$ is of degree p, $\alpha \circ f$ is null-homotopic and extends over Δ'.

Now, let us show that implication (c) implies implication (b). Suppose X is an at most 2-dimensional compactum, A is a closed subset of X, and $g : A \to K(\mathbf{Z}/\mathbf{p}, 1)$ is a map. There is an at most 2-dimensional simplicial complex L (L can be obtained as the nerve of a finite open cover of X of order at most 2), there is a map $\pi : X \to L$, there is a subcomplex K of L, and there is a map $g' : K \to K(\mathbf{Z}/\mathbf{p}, 1)$ so that $\pi(A) \subset K$ and $g' \circ (\pi|A)$ is homotopic to g. Choose an approximate lift $\Pi : X \to P_{\mathbf{Z}/\mathbf{p}}(L)$ of π with respect to π_L. Also, choose an extension $G : P_{\mathbf{Z}/\mathbf{p}}(L) \to K(\mathbf{Z}/\mathbf{p}, 1)$ of $g' \circ \pi_K$. Notice that $G \circ \Pi$ is a homotopy extension of g. Thus, $\dim_{\mathbf{Z}/\mathbf{p}}(X) \leqslant 1$. □

The Edwards–Walsh complex $EW_{\mathbf{Z}}(L, n)$ is constructed by generalizing the procedure of Pontryagin. For simplicity, we fix n and denote $EW_{\mathbf{Z}}(L, n)$ by $EW_{\mathbf{Z}}(L)$. First of all, the n-skeleton of $EW_{\mathbf{Z}}(L)$ is equal to the n-skeleton of L, and π_L is the identity map on the n-skeleton of $EW_{\mathbf{Z}}(L)$. Given an $(n+1)$-simplex Δ of L one creates $EW_{\mathbf{Z}}(\partial\Delta)$ as a $K(\mathbf{Z}, n)$ complex with $\partial\Delta$ being its $(n+1)$-skeleton.

3.7. THEOREM. *Suppose n is an integer. There is a functor $EW_{\mathbf{Z}}$ from SC to CW and there is a natural transformation $\pi_L : EW_{\mathbf{Z}}(L) \to L$ of functors with the following properties:*

(a) *Suppose A is a closed subset of a compactum X so that $\dim_{\mathbf{Z}}(X) \leqslant n$. Given an ordered simplicial complex L, given a map $g : A \to EW_{\mathbf{Z}}(L)$, and given an extension $f : X \to L$ of $\pi_L \circ g$, there is an extension $G : X \to EW_{\mathbf{Z}}(L)$ of g so that G is an approximate lift of f with respect to π_L.*

(b) *Suppose X is a compactum. If any map $f : X \to L$ from X to an ordered simplicial complex L has an approximate lift with respect to π_L, then $\dim_{\mathbf{Z}}(X) \leqslant n$.*

(c) *Given an object L of SC, given a subcomplex K of L, and given a map $f : K \to K(\mathbf{Z}, n)$, the map $f \circ \pi_K$ extends over $EW_{\mathbf{Z}}(L)$.*

(d) *The* $(n+1)$*-skeleton of* $EW_{\mathbf{Z}}(L)$ *coincides with the* n*-skeleton of* $EW_{\mathbf{Z}}(L)$ *and* $\pi_L|EW_{\mathbf{Z}}(L)^{(n)} : EW_{\mathbf{Z}}(L)^{(n)} \to L^{(n)}$ *is an isomorphism of CW complexes.*

SKETCH OF PROOF. If $n < 2$, take $EW_{\mathbf{Z}}(L) = L^{(n)}$ and let π_L be the inclusion. The reason this works is that S^1 is a $K(\mathbf{Z}, 1)$. Thus, assume $n \geqslant 2$. If L is of dimension at most n, then we put $EW_{\mathbf{Z}}(L) = L$ and let $\pi_L : EW_{\mathbf{Z}}(L) \to L$ be the identity map. Now, it suffices to define $EW_{\mathbf{Z}}(\Delta_k)$ for all standard simplices Δ_k with vertices $v_0, v_1, v_2, \ldots, v_k$ so that $v_0 < v_1 < v_2 < \cdots < v_k$ and $k \geqslant n+1$. If $k = n+1$, we attach cells to the n-skeleton of Δ_{n+1} in order to get $EW_{\mathbf{Z}}(\Delta_{n+1})$ as an Eilenberg–MacLane complex $K(\mathbf{Z}, n)$. Obviously, we can do that without attaching cells in dimension $n+1$. Since Δ_{n+1} is contractible, the inclusion $\partial\Delta_{n+1} \to \Delta_{n+1}$ extends to $\pi_L : EW_{\mathbf{Z}}(\Delta_{n+1}) \to \Delta_{n+1}$. Once $EW_{\mathbf{Z}}(\Delta_{n+1})$ is known for all ordered simplices of dimension at most m, one can define $EW_{\mathbf{Z}}(L)$ for all ordered simplicial complexes of dimension at most m. Indeed, suppose L is an ordered simplicial complex of dimension at most m. Notice that $EW_{\mathbf{Z}}(L)$ must contain a copy of $EW_{\mathbf{Z}}(L^{(m-1)})$ as a subcomplex. So it makes sense to build $EW_{\mathbf{Z}}(L)$ by attaching cells to $EW_{\mathbf{Z}}(L^{(m-1)})$. Given an m-simplex s of L, the inclusion $i_s : \partial s \to L^{(m-1)}$ induces an inclusion $j_s : EW_{\mathbf{Z}}(\partial s) \to EW_{\mathbf{Z}}(L^{(m-1)})$. $EW_{\mathbf{Z}}(L)$ is the union of adjunction spaces $EW_{\mathbf{Z}}(L^{(m-1)}) \cup_{j_s} EW_{\mathbf{Z}}(s)$, and $\pi_L : EW_{\mathbf{Z}}(L) \to L$ is the unique extension of $\pi_{L^{(m-1)}} : EW_{\mathbf{Z}}(L^{(m-1)}) \to L^{(m-1)}$ so that $\pi_L|EW_{\mathbf{Z}}(s)$ equals π_s for all s.

Thus, $EW_{\mathbf{Z}}(L)$ is known for all ordered simplicial complexes of dimension at most $n+1$. Suppose that $m > n$, $EW_{\mathbf{Z}}(L)$ is known for all ordered simplicial complexes of dimension at most m, and $EW_{\mathbf{Z}}(\Delta_k)$ is a $K(\mathbf{Z}^{a(k)}, n)$ for each $n < k \leqslant m$. Notice that $EW_{\mathbf{Z}}(\partial\Delta_{m+1})$ is simply connected with the same integral homology (in dimensions from 1 to n) as $\Delta_{m+1}^{(n)}$. As $H_n(\Delta_{m+1}^{(n)})$ is free, it is isomorphic to $\mathbf{Z}^{a(m+1)}$ for some integer $a(m+1)$. Thus, $\pi_n(EW_{\mathbf{Z}}(\partial\Delta_{m+1}))$ is isomorphic to $\mathbf{Z}^{a(m+1)}$ and one can attach cells of dimension at least $n+2$ to $EW_{\mathbf{Z}}(\partial\Delta_{m+1})$ in order to create $EW_{\mathbf{Z}}(\Delta_{m+1})$ as a $K(\mathbf{Z}^{a(m+1)}, n)$. This completes the inductive process of defining $EW_{\mathbf{Z}}(L)$ for all ordered simplicial complexes L.

As in 3.6 one shows condition (a), and one shows that (b) is weaker than (c). Thus, it suffices to prove that (c) is weaker than (d) (the validity of (d) follows from our construction). As in 3.6, it suffices to show this for the case of L being the standard simplex Δ_m. Suppose K is a subcomplex of Δ_m and $f : K \to K(\mathbf{Z}, m)$ is a map. Notice that f can be extended over the n-skeleton of Δ_m. Thus, without loss of generality, it suffices to consider the case $K \supset \Delta_m^{(n)}$. Since there are no $(n+1)$-cells in $EW_{\mathbf{Z}}(L) - EW_{\mathbf{Z}}(K)$, any map $EW_{\mathbf{Z}}(K) \to K(\mathbf{Z}, n)$ extends over $EW_{\mathbf{Z}}(L)$. □

THE IDEA OF THE PROOF OF EDWARDS–WALSH THEOREM 2.6 (*Sketch*). Suppose X is of integral cohomological dimension n, where $n > 1$, and we would like to construct a cell-like map $f : X' \to X$ with X' being of dimension n. A natural guess is to modify $\pi : \widetilde{X}^{(n+1)} \to X$ (see 2.8 and the comment after its proof). The reason is that π is a canonical map which is almost cell-like, namely, it is a UV^n-map. However, $X^{(n+1)}$ is, in general, $(n+1)$-dimensional. The idea is to replace it by an n-dimensional compactum. Consider the projection $X \to X_k$, X_k being the nerve of an open cover of X of mesh at most 2^{-k}. Since X is of integral cohomological dimension n, there is an approximate lift $X \to EW_{\mathbf{Z}}(X_k, n)$ which factors as $X \to X_m \to EW_{\mathbf{Z}}(X_k, n)$ for large m. As you recall, the $(n+1)$-skeleton of $EW_{\mathbf{Z}}(X_k, n)$ coincides with its n-skeleton. Thus, $X_m^{(n+1)} \to X_k^{(n+1)}$ can be approximately factored (or replaced by a map) through an n-dimensional CW complex. This is a rough idea of how to construct a cell-like map. The details are, obviously, more complicated. □

Because of the usefulness of the Edwards–Walsh complexes $EW_{\mathbf{Z}}(L, n)$, one may ponder their existence for groups G other that the integers $\mathbf{Z}$. One may extract the essential features of $EW_{\mathbf{Z}}$ as follows.

3.8. DEFINITION OF EDWARDS–WALSH FUNCTOR EW_G. Suppose n is an integer. A functor $EW_G : \mathcal{SC} \to CW$ is an *Edwards–Walsh functor* provided there is a natural transformation of functors $\pi_L : EW_G(L) \to L$ such that the following conditions are satisfied:

(a) π_L is a combinatorial map and $\pi_L^{-1}(K) = EW_G(K)$ for every subcomplex K of L,
(b) $EW_G(\Delta)$ is a $K(G^m, n)$ for each ordered simplex Δ (here m depends on Δ),
(c) $EW_G(L) = L$ and π_L is the identity if $\dim(L) \leqslant n$, and
(d) if K is a subcomplex of L and $f : K \to K(G, n)$ is a map, then $f \circ \pi_K$ extends over $EW_G(L)$.

The following is the most general result on the existence of Edwards–Walsh functors.

3.9. THEOREM. *An Edwards–Walsh functor exists for each abelian group G such that there is a homomorphism* $\alpha : \mathbf{Z} \to G$ *satisfying the following properties*:

(a) $\alpha \otimes id : \mathbf{Z} \otimes G \to G \otimes G$ *is an isomorphism, and*
(b) $\alpha^* : Hom(G, G) \to Hom(\mathbf{Z}, G)$ *is an isomorphism.*

Theorem 3.9 implies the existence of Edwards–Walsh functors for all Bockstein groups but $\mathbf{Z}/\mathbf{p}^\infty$. In the latter case, an Edwards–Walsh functor fails to exist.

Edwards–Walsh complexes were introduced in [107]. Theorem 3.9 was claimed in [60] under the weaker assumption that α^* is an epimorphism (instead of being an isomorphism). Koyama and Yokoi [77] detected a gap in [60] and provided details that α^* being an isomorphism fills the gap.

4. Dranishnikov realization theorem

Dranishnikov realized the potential of Edwards–Walsh complexes in solving Problems 0.8(b) and 1.9. He used related constructions called *modifications*. We present an outline of Dranishnikov's work using Edwards–Walsh complexes as in [60].

First, let us recall Pontryagin's construction of a 2-dimensional continuum X with $\dim_{\mathbf{Z}/2} X = 1$ (compare with Theorem 0.6).

4.1. EXAMPLE OF X WITH $\dim_{\mathbf{Z}/2}(X) = 1$, $\dim(X) = 2$. *Construction.* X is defined as the inverse limit of the sequence $\cdots X_{k+1} \to X_k \to \cdots$, where $X_1 = S^2$ is the 2-sphere, $X_{k+1} = P_{\mathbf{Z}/2}(X_k)$ and $X_{k+1} \to X_k$ is equal to π_{X_k}, with X_k being considered with a triangulation with simplices being sufficiently small in a given metric on X_k.

Since $\pi_L : P_{\mathbf{Z}/2}(L) \to L$ induces an isomorphism of rational homology, X is of covering dimension 2. It remains to show that $\dim_{\mathbf{Z}/2}(X) = 1$. This property is weaker than the analogous property of the map $\pi_L : P_{\mathbf{Z}/2}(L) \to L$: given a subcomplex K of L and given $f : K \to K(\mathbf{Z}/2, 1)$, the composition $f \circ \pi_K$ extends over $P_{\mathbf{Z}/2}(L)$.

Dranishnikov solved Problem 0.8(b) positively by constructing compacta A_n, $n \geqslant 3$, so that $\dim_{\mathbf{Z}/2}(A_n) = 1$, $\dim(A_n) = n$. The one-point compactification A of the discrete sum of all A_n, $n \geqslant 3$, satisfies $\dim_{\mathbf{Z}/2}(A) = 1$ and $\dim(A) = \infty$. We will outline the construction in the case of $n = 5$.

4.2. EXAMPLE OF X WITH $\dim_{\mathbf{Z}/2}(X) = 1$, $\dim(X) = 5$. *Construction.* A natural attempt is to define X as the inverse limit of the sequence $\cdots X_{k+1} \to X_k \to \cdots$, where $X_1 = S^5$ is the 5-sphere, $X_{k+1} = EW_{\mathbf{Z}/2}(X_k, 1)$ and $X_{k+1} \to X_k$ is equal to π_{X_k}. The major difference between the current construction and Example 4.1 is that $EW_{\mathbf{Z}/2}(X_k, 1)$ is not a finite CW complex. Therefore, it would be impossible to achieve the compactness of X. However, we may try to pick X_{k+1} as a finite subcomplex of $EW_{\mathbf{Z}/2}(X_k, 1)$. This would ensure the compactness of X and, as in Example 4.1, X would be of $\mathbf{Z}/2$-cohomological dimension 1. The only remaining task is to make sure that X is of dimension 5. This is done by keeping $\check{H}^5(X, \mathbf{Q})$ non-zero as follows. The map $\pi_L : EW_{\mathbf{Z}/2}(L, 1) \to L$ induces isomorphisms of all cohomology groups with rational coefficients (this is a combinatorial version of the Vietoris–Begle Theorem – notice that point-inverses of simplices have trivial cohomology groups with rational coefficients). If $c \in H^5(L; \mathbf{Q})$ is non-zero, we choose $c_L \in H^5(EW_{\mathbf{Z}/2}(L, 1); \mathbf{Q})$ such that $c_L = (\pi_L)^*(c)$, and we use the fact that the cohomology theory with rational coefficients is continuous, i.e., there is a finite subcomplex K of $EW_{\mathbf{Z}/2}(L, 1)$ with c_L restricted to K being non-zero. Obviously, we may replace K by its 5-skeleton. Now, use this operation inductively starting with $c_1 \in H^5(S^5; \mathbf{Q})$ being a generator.

By refining the technique of Example 4.2, Dranishnikov was able to solve Problem 1.9 of Bockstein–Boltyanski in positive.

4.3. DRANISHNIKOV'S REALIZATION THEOREM. *For every Bockstein function D, there is a compactum X with $D_X = D$ and* $\dim(X) = \dim(D)$.

One can find the original proof of Theorem 4.3 in [27]. This proof was later modified in [60] by an extensive use of Edwards–Walsh complexes.

5. Infinite-dimensional compacta of finite integral dimension

The next stage in Dranishnikov's program was to solve Problem 0.8(a). It makes sense to mimic Construction 4.2 and attempt to define an infinite-dimensional compactum X of finite integral cohomological dimension as the inverse limit of the sequence $\cdots X_{k+1} \to X_k \to \cdots$, where $X_1 = S^n$ is the n-sphere, $X_{k+1} \subset EW_{\mathbf{Z}}(X_k, m)$ and $X_{k+1} \to X_k$ is equal to $\pi_{X_k} | X_{k+1}$. Since $K(\mathbf{Z}, n)$ has non-trivial cohomology with rational coefficients, one needs to find another continuous cohomology theory which vanishes on $K(\mathbf{Z}, n)$. Luckily, such a cohomology theory already existed in the literature. Namely, the reduced complex K-theory with mod p coefficients vanishes on $K(\mathbf{Z}, 3)$ and, therefore, on $K(\mathbf{Z}^k, 3)$ for all $k \geqslant 1$ (see [4,5]). This leads to a solution of both Alexandroff Problem 0.8(a) and the Cell-like Mapping Problem 2.5 (see [27]):

5.1. DRANISHNIKOV'S THEOREM. *There is an infinite-dimensional compactum X of integral cohomological dimension* 3.

It remained to determine if there is an infinite-dimensional compactum X of integral dimension 2. To be able to mimic Example 4.2, one needed a continuous cohomology theory h^* so that $h^*(K(\mathbf{Z}, 2)) = 0$. While the reduced complex K-theory with $\mathbf{Z/p}$ coefficients of an Eilenberg–MacLane complex $K(\mathbf{Z}, n)$ vanishes for $n \geqslant 3$, the same is not true for $K(\mathbf{Z}, 2)$. (The reader is referred to [4] and [5] for details of these K-theoretic assertions.)

The specific nature of K-theory with $\mathbf{Z/p}$ coefficients ($\mathbf{p}$ a prime) itself plays no direct role in the analysis in [27]. The essential feature is that it is a generalized cohomology theory for which $K(\mathbf{Z}, 3)$ behaves as a point; i.e., the reduced K-theory with $\mathbf{Z/p}$ coefficients of $K(\mathbf{Z}, 3)$ is trivial. The absence of a readily available generalized cohomology theory for which $K(\mathbf{Z}, 2)$ is known to behave as a point requires an alternate approach to that of [27] in order to produce an infinite-dimensional compact metric space having integral cohomological dimension equal to two. In addition, the generalized cohomology theory would have to have the property that, in each dimension, its value on a finite complex is a finite group. The latter is true for K-theory *with finite coefficients*.

If one is asked to create a generalized cohomology theory which helps to detect that a particular compactum X of integral cohomological dimension 2 is of infinite dimension, then one has to connect this cohomology with constructing maps to S^3. Namely, if $\dim_{\mathbf{Z}}(X) = 2$, then $\dim(X) = \infty$ is equivalent to $\dim(X) > 3$, which is equivalent to the existence of a closed subset A of X and a map $f: A \to S^3$ which does not extend over X. Since typical generalized cohomology theories are represented by loop spaces, Dydak and Walsh [58] decided to construct a truncated cohomology theory h^* by defining $h^r(X) = [X, \Omega^{-r} S^3]$ for all $r \leqslant -2$. As shown in [110, p. 125], for any complex L and iterated loop spaces $\Omega^r L$, there is a family of functors $\{X \to [X, \Omega^m L]\}$, associating to a space X the set of homotopy classes of maps of X to $\Omega^m L$. Restricting to $m \geqslant 2$, the functors take values in abelian groups. Unless L happens to be an infinite loop space, these functors are not part of a generalized cohomology theory. Nevertheless, some structure is present. In particular, there is a "truncated" Mayer–Vietoris sequence that can be exploited. To be able to follow the pattern of proofs of 4.2 and 5.1, one needs to verify that $h^*(K(\mathbf{Z}, 2)) = 0$. This is done by observing that $h^*(K(\mathbf{Q}, n)) = 0$ for all $n \geqslant 1$, which amounts to the fact that the higher homotopy groups of S^3 are finite. Now, one converts the inclusion $K(\mathbf{Z}, 2) \to K(\mathbf{Q}, 2)$ to a Serre fibration and one calculates that its fiber is a $K(\mathbf{Q/Z}, 1)$. If $h^*(K(\mathbf{Q/Z}, 1)) = 0$, then we are done. Luckily, the last equality is a consequence of the Sullivan Conjecture proved by Miller [81].

5.2. MILLER'S THEOREM. *If G is a locally finite group* (*i.e., each finitely generated subgroup is finite*) *and L is a finite-dimensional CW-complex, then the space of pointed* *maps from $K(G,1) \to L$, denoted $map_*(K(G,1), L)$, is weakly homotopy equivalent to a point (i.e., has trivial homotopy groups).*

Clearly, $\mathbf{Q}/\mathbf{Z}$ is locally finite. Below are the results from [58] obtained by applying the above method.

5.3. THEOREM. *For $m \geqslant 3$ and an essential map $S^m \to \Omega^3 S^3$, there is a compact metric space X with $\dim_{\mathbf{Z}}(X) \leqslant 2$, and a map $X \to S^m$ such that the composition $X \to S^m \to \Omega^3 S^3$ is essential.*

5.4. COROLLARY. *There exist infinite-dimensional compact metric spaces having integral cohomological dimension equal to two.*

5.5. COROLLARY. *There is a cell-like map $f : \mathbf{R}^5 \to Y$, where $\dim(Y) = \infty$.*

In addition to $\mathbf{Z}$, one can apply this technique to other Bockstein groups, and the next result involves 'mixing' the groups used to measure cohomological dimension. The examples are, in some sense, 'minimally cohomological dimensional' but still infinite dimensional.

5.6. DYDAK–WALSH THEOREM. *There is an infinite-dimensional compactum X with $\dim_{\mathbf{Z}}(X) = 2$, $\dim_{\mathbf{Q}}(X) = 1$, and $\dim_{\mathbf{Z}/\mathbf{p}}(X) = 1$ for each prime $\mathbf{p}$. In particular, $\dim_{\mathbf{Z}}(X \times X) = 3$.*

One can find the details of 5.1 in [27] or [28]. The details of 5.6 can be found in [58] or in [59].

6. Extension types and extension dimension

In [107] the cohomological dimension $\dim_G(X)$ of a compactum was defined as the smallest integer n so that $K(G,n)$ is an absolute extensor of X. In order for this definition to make sense one needs to show that $K(G,n) \in AE(X)$ implies $K(G,n+1) \in AE(X)$ for all integers $n \geqslant 0$. This implication is not a trivial one, and the corresponding result that $S^n \in AE(X)$ implies $S^{n+1} \in AE(X)$ is much easier to prove. These results suggest that perhaps one ought to concentrate on the underlying CW complexes S^n and $K(G,n)$ rather than on the integer n. This was done by Dranishnikov [31], who defined a relation $K \leqslant L$ for CW complexes to mean that for each compact space X, the condition $K \in AE(X)$ implies $L \in AE(X)$. This relation creates an equivalence of CW complexes, and the equivalence class $[K]$ of K is called its *extension type*. The *extension dimension* ext-dim(X), of a compact space X, is defined to be the minimum of extension types of K such that $K \in AE(X)$. Thus, ext-dim$(X) \leqslant K$ means that K is an absolute extensor of X. In particular, ext-dim$(X) \leqslant S^n$ is equivalent to the covering dimension of X being at most n, and ext-dim$(X) \leqslant K(G,n)$ is equivalent to the cohomological dimension of X being at most n. In this way, the extension dimension generalizes both the covering dimension and the cohomological dimension. The cornerstone of Dranishnikov's generalized dimension theory is the following.

6.1. DRANISHNIKOV'S DUALITY THEOREM. *The extension dimension of a compact space exists, and each extension type is equal to the extension dimension of some compact space.*

This theorem shows that there is a duality between CW complexes and compact spaces. There is a natural question of characterizing compact spaces of the same extension dimension, and [31] contains a partial answer to it.

6.2. THEOREM. *If X and Y are finite-dimensional compact spaces, then the extension dimensions of $X \times I$ and $Y \times I$ coincide if and only if the dimension functions of X and Y coincide.*

Dydak and Dranishnikov extended the theory of extension dimension to a larger class of spaces (so-called *ext-spaces*) which includes all compact spaces and all separable metrizable spaces. Also, they consider the dual relation $X \leqslant Y$ for ext-spaces which means that $K \in AE(Y)$ implies $K \in AE(X)$ for all CW complexes K. This relation creates an equivalence relation of ext-spaces, and the equivalence class of X is called its *dimension type* and denoted by $[X]$. Given a CW complex K one considers the family of all dimension types $[X]$ such that $K \in AE(X)$. If this family has a maximum, then it is called the *extension universe* of K and is denoted by ext-univ(K). It is shown that there is a formal duality between the concept of extension dimension and the concept of extension universe. Also, there is an informal duality between results dealing with extension types and results dealing with dimension types. Some pieces of this duality are still unknown and have been posed as open problems.

Extension dimension deals with general topological spaces and, in the case of finite-dimensional compacta, is related to the Bockstein algebra as seen in 6.2. There is a theory dual to extension dimension which deals with CW complexes and, in the case of countable CW complexes, one has an associated algebraic object called the *dual Bockstein algebra*.

The whole program can be carried out for a subclass of CW complexes. The main example is the class of wedges of Eilenberg–MacLane complexes. It is shown that in this case, the major results of the cohomological dimension theory correspond to the fundamental questions in the whole program. Namely, the existence of the extension dimension corresponds to the First Bockstein Theorem and the existence of the extension universe corresponds to the Dranishnikov Realization Theorem. The Second Bockstein Theorem corresponds to the existence of the Cartesian product of dimension types and its dual corresponds to the existence of the smash product of extension types.

In the remainder of this section we describe the results of [38] and [48].

In the classical definition of dimension one initially defines the relation $\dim(X) \leqslant n$, which is then used to define $\dim(X)$. In the same spirit, one may talk about the relation ext-dim$(X) \leqslant M$ without discussing the existence of ext-dim(X).

6.3. DEFINITION OF ext-dim$(X) \leqslant M$. We say that the *extension dimension* of a topological space X is less than or equal to M (notation: ext-dim$(X) \leqslant M$), provided the property $M \in AE(X)$ holds.

6.4. EXAMPLE. If S^n is the n-dimensional sphere and X is a normal space, then ext-dim$(X) \leqslant S^n$ means $\dim X \leqslant n$.

Traditionally, for a normal space X, one declares $\dim(X)$ to be infinite if $\dim(X) \leqslant n$ does not hold for any integer n. From the point of view of extension theory it is more natural to interpret $\dim(X) \leqslant \infty$ as $S^\infty \in AE(X)$, where $S^\infty = \bigcup_{n=0}^{\infty} S^n$. S^∞ is a contractible CW complex and one finds it natural to consider spaces X so that any contractible CW complex K is an absolute extensor of X.

6.5. DEFINITION OF A CW-SPACE. A space X (not, in general, required to be Hausdorff) is called a *cw-space* provided it is a k-space and $K \in AE(X)$ for all contractible CW complexes K.

Recall that X is a k-space if A is closed in X if and only if $A \cap Z$ is closed for all compact subsets Z of X.

As we mentioned, at certain moment a natural question of the existence of the extension dimension pops up. It turns out that one can prove its existence for ext-spaces.

6.6. DEFINITION OF AN EXT-SPACE. A cw-space X is called an *ext-space* if for any CW complex $K \in AE(X)$, and any subcomplex L of K containing at most $w(X)$ cells ($w(X)$ is the weight of X) there is a subcomplex M containing at most $w(X)$ cells such that $M \in AE(X)$ and $L \subset M$.

Recall that the *weight* of a topological space X is the minimal cardinal number of elements in a basis of X.

6.7. THEOREM. *The class of ext-spaces contains all compact Hausdorff spaces and all separable metrizable spaces.*

6.8. NOTATION. The class of all *CW complexes* is denoted by $\mathcal{CW}$.

The class of all *cw-spaces* is denoted by $\mathcal{CW}$-$\mathcal{SPACES}$.

The class of all *ext-spaces* is denoted by $\mathcal{EXT}$.

The class of all *compact Hausdorff spaces* is denoted by $\mathcal{COMPACT}$.

The class of all *compact metrizable spaces* is denoted by $\mathcal{COMPACTA}$.

The class of all *finite-dimensional compact metrizable spaces* is denoted by $\mathcal{FD}$-$\mathcal{COMPACTA}$.

The class of all *separable metrizable spaces* is denoted by $\mathcal{SEPARABLE}$.

6.9. DEFINITION OF EXTENSION TYPES AND DIMENSION TYPES. Suppose $\mathcal{C}$ is a class of topological spaces and $\mathcal{M} \subset \mathcal{CW}$. For simplicity we ignore the possibility that either $\mathcal{C}$ or $\mathcal{M}$ may not be a set.

We define a relation $\leqslant_{\mathcal{C}}$ on $\mathcal{M}$ as follows: $K \leqslant_{\mathcal{C}} L$ for $K, L \in \mathcal{M}$ if and only if $K \in AE(X)$ implies $L \in AE(X)$ for every $X \in \mathcal{C}$. Then, we define an equivalence relation $K \sim_{\mathcal{C}} L$ if both $K \leqslant_{\mathcal{C}} L$ and $L \leqslant_{\mathcal{C}} K$ hold. An equivalence class under this relation is called an *extension type* or, more precisely, $(\mathcal{C}, \mathcal{M})$-*extension type*. K is said to be of *trivial extension* *type* if $X \tau K$ for all $X \in \mathcal{C}$. The set of all extension types is denoted by ExtTypes($\mathcal{C}, \mathcal{M}$). Clearly $\leqslant_{\mathcal{C}}$ induces a partial order on ExtTypes($\mathcal{C}, \mathcal{M}$). If it is not ambiguous, we denote our partial order simply by $\leqslant$.

Dually, we define a relation $\leqslant_{\mathcal{M}}$ on $\mathcal{C}$ as follows: $X \leqslant_{\mathcal{M}} Y$ for $X, Y \in \mathcal{C}$ if and only if ext-dim$(Y) \leqslant K$ implies ext-dim$(X) \leqslant K$ for every $K \in \mathcal{M}$. Then, we define an equivalence relation $X \sim_{\mathcal{M}} Y$ if both $X \leqslant_{\mathcal{M}} Y$ and $X \leqslant_{\mathcal{M}} Y$ hold. An equivalence class under this relation is called a *dimension type* or, more precisely, $(\mathcal{C}, \mathcal{M})$*-dimension type*. X is said to be of *trivial dimension type* if $X \tau K$ for all $K \in \mathcal{M}$. The set of all dimension types is denoted by DimTypes($\mathcal{C}, \mathcal{M}$). Clearly $\leqslant_{\mathcal{M}}$ induces a partial order on DimTypes($\mathcal{C}, \mathcal{M}$). If it is not ambiguous, we denote our partial order simply by $\leqslant$.

6.10. EXAMPLES. Suppose $\mathcal{S}$ is the set of spheres of all dimensions (including infinity).

(1) $S^n \leqslant_{\mathcal{C}} S^m$ for any class $\mathcal{C}$ if $n \leqslant m < \infty$,

(2) if X, Y are cw-spaces, then $X \leqslant_{\mathcal{S}} Y$ if and only if $\dim(X) \leqslant \dim(Y)$,

(3) let $\mathcal{C} = \mathcal{COMPACTA}$ and $n > 1$. Then $S^n \leqslant_{\mathcal{C}} K(\mathbf{Z}, n)$ but $S^n \sim_{\mathcal{C}} K(\mathbf{Z}, n)$ does not hold.

6.11. NOTATION. For every $M \in \mathcal{M}$, the corresponding class in ExtTypes($\mathcal{C}, \mathcal{M}$) will be denoted by $[M]$, or simply by M if the latter is not ambiguous. Dually, for every $X \in \mathcal{C}$, the corresponding class in DimTypes($\mathcal{C}, \mathcal{M}$) will be denoted by $[X]$, or simply by X if the latter is not ambiguous.

Notice that there are minimal and the maximal elements in ExtTypes($\mathcal{CW}$-$\mathcal{SPACES}$, $\mathcal{CW}$). The minimal element is generated by the two-points complex D and the maximal element is generated by the one point complex P. Given a space X we consider the class

$$\{[M] \in \text{ExtTypes}(\mathcal{C}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}.$$

If this class has a minimal element $[K]$, then K is one of the best approximations of X in $\mathcal{M}$ from the point of view of extension theory.

6.12. DEFINITION OF EXTENSION DIMENSION AND EXTENSION UNIVERSE. Suppose $\mathcal{M} \subset \mathcal{CW}$, $\mathcal{C} \subset \mathcal{CW}$-$\mathcal{SPACES}$ and X is a space. If

$$\{[M] \in \text{ExtTypes}(\mathcal{C}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}$$

has a minimal element, then it is called an *extension dimension* and denoted by ext-dim$_{(\mathcal{C}, \mathcal{M})}(X)$.

Notice that this notation agrees with the previous notation ext-dim$(X) \leqslant M$. Moreover, if X is an ext-space, then ext-dim(X) may be interpreted as

$$\text{ext-dim}_{(\mathcal{CW}\text{-}\mathcal{SPACES}, \mathcal{CW})}(X).$$

Dually, if

$$\{[X] \in \text{DimTypes}(\mathcal{C}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}$$

has a maximal element, then it is called an *extension universe* and denoted by ext-univ$_{(\mathcal{C},\mathcal{M})}(M)$.

The following proposition explains our definition of the dimension type.

6.13. PROPOSITION. *Let $\mathcal{C} = \mathcal{CW}\text{-}\mathcal{SPACES}$ and let $\mathcal{M}$ be the set of spheres of all dimensions (including infinity). Then*

(1) *the dimension type $[X]$ consists of all spaces Y of the same dimension as X,*
(2) *the extension dimension of X is the sphere $S^{\dim X}$, and*
(3) *the extension universe of S^n is $[X]$ for any space X of dimension n.*

The following result explains the duality between the concept of extension dimension and the concept of extension universe.

6.14. DUALITY THEOREM.

(1) *Let $X \in \mathcal{C}$. If* ext-dim$_{(\mathcal{C},\mathcal{M})}(X)$ *exists and equals $[K]$, then* ext-univ$_{(\mathcal{C},\mathcal{M})}(K)$ *exists and equals $[X]$.*
(2) *Let $K \in \mathcal{M}$. If* ext-univ$_{(\mathcal{C},\mathcal{M})}(K)$ *exists and equals $[X]$, then* ext-dim$_{(\mathcal{C},\mathcal{M})}(X)$ *exists and equals $[K]$.*

The following is the most general result regarding the existence of extension dimension.

6.15. THEOREM. *Suppose $\mathcal{M} \subset \mathcal{CW}$ is closed under taking arbitrary wedges and under taking subcomplexes. If $\mathcal{C} \subset \mathcal{CW}\text{-}\mathcal{SPACES}$, then for every ext-space X the class*

$$\{[M] \in \text{ExtTypes}(\mathcal{C}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}$$

has a minimal element.

The subsequent two results deal with the existence of an extension universe.

6.16. THEOREM. *For each CW complex K, there is a compact space X such that $[X] =$ ext-univ$_{(\mathcal{COMPACT},\mathcal{CW})}(K)$. Moreover, if K is homotopy dominated by a finite CW complex, then X is compact metric.*

6.17. THEOREM. *For every countable CW complex $K \in \mathcal{M}$, there is a separable metrizable space X such that*

$$[K] = \text{ext-dim}_{(\mathcal{SEPARABLE},\mathcal{M})}(X).$$

Besides the formal duality between extension dimension and extension universe, there is an informal duality between extension types and dimension types. In Table 1 there are some dual statements.

For further information on the topics of cohomological and extension dimension, the reader should consult A. Chigogidze's article [19] in this Handbook.

Table 1

Compact spaces $\mathcal{C}$	CW complexes $\mathcal{M}$
The inverse limit of at most K-dimensional compact spaces is at most K-dimensional, where $K \in \mathcal{M}$.	The direct limit of CW complexes which are greater than or equal to a given extension type is greater than or equal to that type.
Every at most K-dimensional compact space can be presented as the limit space of a σ-system of at most K-dimensional compacta.	If ext-dim$(X) \leqslant K$, then K is the direct limit of $w(X)$-complexes K_α with ext-dim$(X) \leqslant K_\alpha$.
The function ext-dim is well-defined.	The function ext-dim is surjective.
Countability Problem: If X is metrizable, can ext-dim(X) be represented by a countable complex?	Metrizability Problem: If M is countable, does there exist a metrizable X with ext-dim$(X) = M$?
$X \in ANR$	?
$\dim X = n$	$K = S^n$
$\dim_G X = n$	$K = K(G, n)$

6.1. *Classical cohomological dimension theory reinterpreted*

The purpose of this section is to show how the basic results of classical cohomological dimension theory can be interpreted using the concepts of extension type, dimension type, extension dimension, and the extension universe. Typically, there are two interpretations: one in terms of compact spaces and the other in terms of CW complexes.

6.18. DEFINITION. Let $\mathcal{EM}$ be the class of all possible wedges of Eilenberg–MacLane spaces $K(G,n)$, where $n \in \mathbf{Z}_+ \cup \{\infty\}$ and G is an abelian group (if $n = \infty$, we define $K(G,n)$ to be the one-point space).

The following is a restatement of the First Bockstein Theorem in terms of CW complexes.

6.19. THEOREM. *$K(G,n)$ is of the same extension type as*

$$\bigvee_{H \in \sigma(G)} K(H,n)$$

in ExtTypes$(\mathcal{COMPACT}, \mathcal{EM})$.

Here is an interpretation of the First Bockstein Theorem in terms of compact spaces.

6.20. THEOREM. *Suppose $X, Y \in \mathcal{C} = \mathcal{COMPACT}$. Then,*

(1) *$X \leqslant_{\mathcal{EM}} Y$ if and only if $D_X \leqslant D_Y$, $X \sim_{\mathcal{EM}} Y$ if and only if $D_X = D_Y$, and*

(2) *the extension dimension* ext-dim$_{(\mathcal{C},\mathcal{EM})}(X)$ *of* X *is*

$$\bigvee_{H\in\sigma} K(H, \dim_H X).$$

Here is an interpretation of Bockstein functions in terms of CW complexes.

6.21. THEOREM. *Let* $\mathcal{C} = \mathcal{COMPACT}$. *Given* $K \in \mathcal{EM}$, *consider the set* $\mathcal{F}$ *of functions* D *such that* $K \sim_{\mathcal{C}} \bigvee_{H\in\sigma} K(H, D(H))$. *Then*
(1) $\mathcal{F} \neq \emptyset$, *and*
(2) $\inf \mathcal{F}$ *belongs to* $\mathcal{F}$ *and is a Bockstein function.*

The next result is an interpretation of (and is equivalent to) the Dranishnikov Realization Theorem.

6.22. THEOREM. *Let* $\mathcal{C} = \mathcal{COMPACT}$. *Then*
(1) *any* $K \in \mathcal{EM}$ *is equivalent to the unique* $\bigvee_{H\in\sigma} K(H, D(H))$, *where* D *is a Bockstein function, and*
(2) *any* $K \in \mathcal{EM}$ *has its extension universe which is a compactum.*

6.2. *Dual Bockstein algebra*

Our next goal is to define Bockstein functions of CW complexes as the dual notion to Bockstein functions of compact spaces. In view of Theorem 6.22, one can define Bockstein functions of elements of $\mathcal{EM}$ in a convoluted way. We plan to define these for all CW complexes in a way dictated by duality. Since $D_X(H)$ is defined as the minimum n with ext-dim$(X) \leqslant K(H, n)$, the following definition makes sense.

6.23. DEFINITION OF EXTENSION FUNCTION OF A CW COMPLEX. For any CW complex L its *extension function* $E^L : \sigma \to \mathbf{Z}_+ \cup \{\infty\}$ is defined as follows:

$$E^L(H) = \min\{n \in \mathbf{Z}_+ \cup \{\infty\} \mid L \leqslant_{\mathcal{C}} K(H, n)\},$$

where $\mathcal{C} = \mathcal{FD}\text{-}\mathcal{COMPACTA}$ is the class of finite-dimensional compacta. Notice that the extension function of L depends on its extension type only.

Let us explain the above definition. First, let us define $K(D)$ for any function D.

6.24. DEFINITION OF $K(D)$. If D is a function on Bockstein groups, then $K(D)$ is defined as

$$\bigvee_{H\in\sigma} K(H, D(H)).$$

Here is the meaning of Definition 6.23: Consider the set $\mathcal{F}$ of all functions D such that $L \leqslant_{\mathcal{C}} K(D)$. Notice that $D \equiv \infty$ belongs to $\mathcal{F}$ and $\inf \mathcal{F} \in \mathcal{F}$. As in 6.21, $\inf \mathcal{F}$ is a Bockstein function and it is clear that it is equal to E^L.

6.25. PROPOSITION. *Let $\mathcal{C} = \mathcal{FD}$-$\mathcal{COMPACTA}$.*
(1) *If L is simply connected, then $L \sim_{\mathcal{C}} \bigvee_{H \in \sigma} K(H, E^L(H))$.*
(2) *If L is a countable CW complex, then $E^{\Sigma L} = E^L + 1$.*

Extension functions are numerical objects which are useful in discussing the existence of extension dimension.

6.26. PROPOSITION. *Let $X \in \mathcal{C} = \mathcal{FD}$-$\mathcal{COMPACTA}$ and let K be a CW complex.*
(1) *If $X \tau K$, then $D_X \leqslant E^K$.*
(2) *If K is simply connected, then $X \tau K$ if and only if $D_X \leqslant E^K$.*
(3) *If $\mathcal{M} \subset \mathcal{CW}$ contains all countable Eilenberg–MacLane complexes and $[K] =$ ext-dim$_{(\mathcal{C}.\mathcal{M})}(X)$, then $E^K = D_X$.*
(4) *If $K \leqslant L$, then $E^K \leqslant E^L$.*
(5) *If $L \in \mathcal{CW}$ is simply connected, then $K \leqslant L$ if and only if $E^K \leqslant E^L$.*
(6) *If D is a Bockstein function and $L = K(D)$, then $E^L = D$.*

One may view the definition of the wedge of Bockstein functions (see Definition 1.11) as the one that assures the identity $D_{X \vee Y} = D_X \vee D_Y$. From the point of view of CW complexes, however, one would like to achieve the identity $E^{K \vee L} = E^K \vee E^L$. This forces us to introduce a new wedge operation for Bockstein functions, namely the cw-wedge of Bockstein functions. In [38] one has a formal definition of the cw-wedge. Here we give a more geometric interpretation of this definition.

6.27. DEFINITION OF THE CW-WEDGE OF BOCKSTEIN FUNCTIONS. Given a set $\{D_a\}_{a \in A}$ of Bockstein functions, one defines its *cw-wedge* $\bigvee^{cw}_{a \in A} D_a$ as the extension function of the wedge $\bigvee_{a \in A} K(D_a)$.

Let us explain the sense in which the cw-wedge of Bockstein functions is dual to the wedge of Bockstein functions.

6.28. PROPOSITION. *Let $\mathcal{BF}$ be the set of all Bockstein functions and let $\{D_a\}_{a \in A} \subset \mathcal{BF}$. Then,*
(1) $\bigvee_{a \in A} D_a = \inf\{E \in \mathcal{BF} \mid E \geqslant D_a$ *for all* $a \in A\}$, *and*
(2) $\bigvee^{cw}_{a \in A} D_a = \sup\{E \in \mathcal{BF} \mid E \leqslant D_a$ *for all* $a \in A\}$.

The following is the dual to 1.14(1).

6.29. THEOREM. *If $L = \bigvee_{i=1}^{\infty} L_i$ is a countable CW complex, then*

$$E^L = \bigvee_{i \geqslant 1}^{cw} E^{L_i}.$$

In [38] there is a formal definition of the smash product of Bockstein functions. Let us define it in a more geometric way.

6.30. DEFINITION OF THE SMASH PRODUCT OF BOCKSTEIN FUNCTIONS. Suppose D_1 and D_2 are two Bockstein functions. Define the *smash product* $D_1 \wedge D_2$ as the extension function of $K(D_1) \wedge K(D_2)$.

Kuzminov [80] found a basis (called the *Kuzminov basis*) of the Bockstein algebra $(\mathcal{BF}, \vee, \times)$. It is a countable subset $\mathcal{K}$ of the set of Bockstein functions $\mathcal{BF}$ such that any element D of $\mathcal{BF}$ can be expressed by elements of $\mathcal{K}$ using the operations $\times$ and $\vee$. The first proof of the Dranishnikov Realization Theorem [27] amounted to constructing a compactum X_K for each $K \in \mathcal{K}$ so that $D_{X_K} = K$ and $\dim(X_K) = \dim(K)$. We plan to introduce a basis of the Dual Bockstein Algebra $(\mathcal{BF}, \bigvee^{cw}, \wedge)$. In a sense, our basis is simpler than the Kuzminov basis and is more intuitive. Elements of our basis are defined as extension functions of Eilenberg–MacLane complexes and the Kuzminov basis can be viewed as its dual.

6.31. DEFINITION. For each $H \in \sigma$ and each $n > 0$, let $E^{H,n} = E^{K(H,n)}$. Let $E^H = E^{H,1}$.

The following is the dual to 1.14(2).

6.32. THEOREM. *If K and L are two countable CW complexes, then*

$$E^{K \wedge L} = E^K \wedge E^L.$$

6.33. THEOREM. *$\{0\} \cup \{E^H\}_{H \in \sigma}$ is a basis of the dual Bockstein algebra.*

The following is an interpretation of the Second Bockstein Theorem:

6.34. THEOREM. *Suppose $\mathcal{C}$ is a class of finite-dimensional compacta and $\mathcal{M}$ is a class of CW complexes containing all countable Eilenberg–MacLane complexes. Then*

(1) *if $\mathcal{C}$ is closed under the Cartesian product, then the operation $[X] \times [Y] = [X \times Y]$ is well-defined in* DimTypes$(\mathcal{C}, \mathcal{M})$, *and*

(2) *if $\mathcal{C}$ is closed under the smash product, then the operation $[X] \wedge [Y] = [X \wedge Y]$ is well-defined on non-trivial dimension types in* DimTypes$(\mathcal{C}, \mathcal{M})$.

The following presents a situation in which one can compute the extension dimension.

6.35. THEOREM. *Let $\mathcal{C} = \mathcal{FD\text{-}COMPACTA}$ and let $\mathcal{M}$ be a class of CW complexes containing all countable CW complexes. Then*

(1) *if $X, Y \in \mathcal{C}$ are of non-zero dimension and $Z = X \times Y$, then*

$$\text{ext-dim}_{(\mathcal{C},\mathcal{M})}(Z) = \left[\bigvee_{H \in \sigma} K(H, D_Z(H)) \right],$$

and

(2) *if* $X, Y \in \mathcal{C}$ *are of non-zero dimension and* $\infty > \dim Y = \dim_H Y$ *for all* $H \in \sigma$, *then,*

$$\text{ext-dim}_{(\mathcal{C},\mathcal{M})}(X \times Y) = \text{ext-dim}_{(\mathcal{C},\mathcal{M})}(X) \wedge \text{ext-dim}_{(\mathcal{C},\mathcal{M})}(Y).$$

Finally, we present the situation in which having the same extension dimension is equivalent to having the same cohomological dimension (the case $Z = I$ is in 6.2).

6.36. COROLLARY. *Let* $\mathcal{C} = \mathcal{FD}\text{-}\mathcal{COMPACTA}$ *and let* $\mathcal{M}$ *be a class of CW complexes containing all countable CW complexes. Suppose Z is a continuum such that* $\infty > \dim Z = \dim_H Z$ *for all* $H \in \sigma$. *The following two properties are equivalent*:

(1) $\text{ext-dim}_{(\mathcal{C},\mathcal{M})}(X \times Z) = \text{ext-dim}_{(\mathcal{C},\mathcal{M})}(Y \times Z)$ *in* $\text{ExtTypes}(\mathcal{C}, \mathcal{M})$.

(2) $\dim_G X = \dim_G Y$ *for all abelian groups G.*

An explanation of the above result is contained in [48].

6.37. LEMMA. *Let* $\mathcal{C} = \mathcal{FD}\text{-}\mathcal{COMPACTA}$ *and let* $\mathcal{M}$ *be be a class of CW complexes containing all countable CW complexes. Suppose* $K \leqslant_{\mathcal{C}} L$. *Then*

(a) *if K is simply connected, then so is L, and*

(b) *if K is homologically n-connected (i.e.,* $\widetilde{H}_i(K; \mathbf{Z}) = 0$ *for* $i \leqslant n$*), then so is L.*

The above lemma implies that the following definition makes sense.

6.38. DEFINITION OF SIMPLY CONNECTED EXTENSION TYPE. We say that the extension dimension of a finite-dimensional compactum X is *simply connected* (respectively, *homologically n-connected*) provided K is simply connected (respectively, homologically n-connected) for any CW complex K which is an absolute extensor of X.

Here is an explanation of Corollary 6.36 (see [48]).

6.39. THEOREM. *Let* $\mathcal{C} = \mathcal{FD}\text{-}\mathcal{COMPACTA}$ *and let* $\mathcal{M}$ *be a class of CW complexes containing all countable CW complexes. If X and Y are two elements of* $\mathcal{C}$ *whose extension dimensions are simply connected, then* $\text{ext-dim}_{(\mathcal{C},\mathcal{M})}(X) = \text{ext-dim}_{(\mathcal{C},\mathcal{M})}(Y)$ *in* $\text{ExtTypes}(\mathcal{C}, \mathcal{M})$ *if and only if* $D_X = D_Y$.

6.40. THEOREM. *The extension dimension of a compactum X is homologically n-connected if and only if* $D_X \geqslant n + 1$.

6.41. DEFINITION OF DIMENSION ALGEBRA AND EXTENSION ALGEBRA. Let $\mathcal{C} = \mathcal{FD}\text{-}\mathcal{COMPACTA}$ and let $\mathcal{M}$ be the class of all countable CW complexes. The *dimension algebra* is $(\text{DimTypes}(\mathcal{C}, \mathcal{M}), \vee, \times)$ and the *extension algebra* is $(\text{ExtTypes}(\mathcal{C}, \mathcal{M}), \vee, \wedge)$.

6.42. THEOREM. *There is a natural, non-injective homomorphism*

$$\Phi : \big(\text{DimTypes}(\mathcal{C}, \mathcal{M}), \vee, \times\big) \to (\mathcal{BF}, \vee, \times)$$

from the dimension algebra to the Bockstein algebra.

6.43. THEOREM. *There is a natural, non-injective homomorphism*

$$\Psi : \big(\mathrm{ExtTypes}(\mathcal{C}, \mathcal{M}), \vee, \wedge\big) \to \left(\mathcal{BF}, \bigvee^{cw}, \wedge\right)$$

from the extension algebra to the dual Bockstein algebra.

Proofs of Theorems 6.1 and 6.2 can be found in [31]. All of the remaining results of this section (with the exception of 6.37–6.40) can be found in [38]. Proofs of 6.37–6.40 will be published soon in [48].

7. Extension theory

In 1992, during a workshop on cohomological dimension theory in Knoxville, TN, A. Dranishnikov presented certain results (see [31] or 6.1–6.2 in this paper) of generalized dimension theory which he called "extension theory". Inspired by his talk, the author thought that perhaps extension theory could be a theory encompassing both the covering and the cohomological dimension theories. Up to that point results and conjectures in cohomological dimension theory were created by formally replacing dim with $\dim_G$. The author's idea was that one ought to create results/conjectures which would formally imply the corresponding results in both dimension theory and cohomological dimension theory. Let us demonstrate this on the following basic example: Since at that time the author was working on Kuzminov's question "Does $\dim_G(A \cup B) \leqslant \dim_G(A) + \dim_G(B) + 1$ hold?", which was formally created from the Urysohn–Menger theorem which states that $\dim(A \cup B) \leqslant \dim(A) + \dim(B) + 1$, it was natural to try to translate the inequality

$$\dim(A \cup B) \leqslant \dim A + \dim B + 1$$

into the language of absolute extensors. A simple translation is:

$$\text{if } S^m \in AE(A) \text{ and } S^n \in AE(B), \text{ then } S^{m+n+1} \in AE(A \cup B).$$

What is the connection between S^m, S^n, and S^{m+n+1}?

The answer is: S^{m+n+1} is the join of S^m and S^n.

Using this observation, the author [47] proposed the following Conjectures 7.1–7.7 (either as specific conjectures or problems which subsequently were converted into conjectures in [53]).

7.1. CONJECTURE. If $K \in AE(A)$ and $L \in AE(B)$ are CW complexes, then $K * L \in AE(A \cup B)$.

7.2. CONJECTURE. If $K * L \in AE(X)$, then there are subsets A, B of X such that $X = A \cup B$ and $K \in AE(A)$ and $L \in AE(B)$.

7.3. CONJECTURE. If $K \in AE(X)$ is a countable CW complex, then there is a completion $\overline{X}$ of X with $K \in AE(\overline{X})$.

7.4. CONJECTURE. Let K be a countable CW complex. Any X with $K \in AE(X)$ admits compactification $\overline{X}$ with $K \in AE(\overline{X})$ if and only if K is homotopy dominated by a finite polyhedron.

7.5. CONJECTURE. If K, L are CW complexes and $K \in AE(X)$, $L \in AE(Y)$, then $K \wedge L \in AE(X \times Y)$.

7.6. CONJECTURE. Suppose K is a countable CW complex. There is a universal separable space in the class $\{X \mid K \in AE(X)\}$.

7.7. CONJECTURE. Suppose K is a CW complex. $\{X \mid K \in AE(X)\}$ has a universal compact space if and only if K is homotopy dominated by a finite polyhedron.

Several of the conjectures above were subsequently answered in the positive as indicated by the following list.

(a) Conjecture 7.1 by the author [47] for arbitrary metrizable spaces A, B and arbitrary CW complexes K, L.
(b) Conjecture 7.2 by Dranishnikov [34] for compact spaces X and by Dranishnikov and Dydak [39] for separable metric spaces.
(c) Conjecture 7.3 by Olszewski [89].
(d) Conjecture 7.5 by Dranishnikov and Dydak [39] in the case of X and Y being finite-dimensional, and Y being compact.
(e) Conjecture 7.6 by Olszewski [90].

Not only do extension theory results have simpler proofs than their predecessors in cohomological dimension theory, but they also explain certain phenomena in cohomological dimension theory and are sometimes more general than expected. Let us demonstrate this in the case of a result of Dranishnikov [34] result:

7.8. THEOREM. *Suppose X is a compactum and K, L are countable CW complexes. If $K * L \in AE(X)$, then there are subsets A, B of X such that $X = A \cup B$, $K \in AE(A)$, and $L \in AE(B)$.*

He applied this result successfully to the Mapping Intersection Problem in codimension different from two. It turns out that Theorem 7.8 also implies Dranishnikov's Realization Theorem.

In the process of proving Theorem 7.8, Dranishnikov generalized a useful result from classical dimension theory.

7.9. EILENBERG–BORSUK THEOREM. *If X is a separable metrizable space of dimension n, then for any map $f: A \to S^k$, A closed in X and $k < n$, there is an extension $f': U \to K$ of f over an open set U such that* $\dim(X - U) \leqslant n - k - 1$.

Here is Dranishnikov's generalization (see [34]).

7.10. GENERALIZED EILENBERG THEOREM. *Let K, L be countable CW complexes. If X is a compactum and $K * L$ is an absolute extensor of X, then for any map $f : A \to K$, A closed in X, there is an extension $f' : U \to K$ of f over an open set U such that $L \in AE(X - U)$.*

This result was verified for X separable and metrizable by Dranishnikov and Dydak [39].

7.11. GENERALIZED EILENBERG–BORSUK THEOREM. *Let L be a countable CW complex. If X is a separable metrizable space and $K * L$ is an absolute extensor of X for some CW complex K, then for any map $f : A \to K$, A closed in X, there is an extension $f' : U \to K$ of f over an open set U such that $L \in AE(X - U)$.*

[39] also contains extension theory version of the following classical result.

7.12. THEOREM. *If X is a separable metrizable space of finite dimension n, then for any $k < n$ there is a closed subset Y of X with* $\dim(Y) = k$.

Here is the corresponding result in cohomological dimension theory.

7.13. THEOREM. *Suppose $G_i, \ldots, G_n$ are countable, non-trivial, abelian groups and $k > 0$. For any separable metrizable space X of finite positive dimension, there is a closed subset Y of X with* $\dim_{G_i}(Y) = \max(\dim_{G_i}(X) - k, 1)$ *for* $i = 1, \ldots, n$.

When translating results from extension theory to cohomological dimension theory we find that the following theorem is of fundamental importance.

7.14. THEOREM. *Suppose X is a metrizable space and K is a connected CW complex. Consider the following conditions*:

(1) $K \in AE(X)$.
(2) $SP^{\infty}(K) \in AE(X)$.
(3) $\dim_{H_m(K;\mathbf{Z})}(X) \leqslant m$ *for all* $m \geqslant 0$.
(4) $\dim_{\pi_m(K;\mathbf{Z})}(X) \leqslant m$ *for all* $m \geqslant 0$.

Then, (1) *implies* (2). *If K is simply connected, then conditions* (2)–(4) *are equivalent. If X is of finite dimension and K is simply connected, then conditions* (1)–(4) *are equivalent.*

Recall that $SP^{\infty}(K)$ is the infinite symmetric product of K (see [26]).

For X compact, Theorem 7.14 is due to Dranishnikov [31]. Subsequently, it was generalized to metrizable spaces by Dydak [45,47].

Theorem 7.14 is reminiscent of the classical Hurewicz Theorem, and in fact part of its proof relies on the Serre version of the Hurewicz Theorem. Thus, it represents a point of overlap between extension theory and algebraic topology. As we mentioned, cell-like maps represent a point of overlap between extension theory an geometric topology. It seems natural to investigate points of overlap of extension theory and point set topology. This was done by the author [49] and the next part of this section is devoted to an account of [49].

One of the most beautiful results of general topology is:

7.15. TIETZE–URYSOHN THEOREM. *Given a closed subset A of a normal space X and given $f : A \to [0, 1]$, f extends over X.*

The modern way of stating Tietze–Urysohn Theorem is:

7.15′. THEOREM. $[0, 1]$ *is an absolute extensor of all normal spaces.*

A big part of the Polish School of Topology was the study of absolute extensors for metrizable spaces, the so-called AR's. In general topology a dual approach was undertaken: given a space (or a class of spaces) K, characterize spaces X such that $K \in AE(X)$.

In the process of proving the Tietze–Urysohn Theorem one needs:

7.16. URYSOHN LEMMA. *Given a closed subset A of a normal space X, and given $f : A \to S^0 \subset I$, f extends over X to $F : X \to I$.*

7.16 means that one ought to generalize the concept of an absolute extensor to pairs of topological spaces.

7.17. DEFINITION. Let A be a subspace of X and let L be a subspace of K. Then $(K, L) \in AE(X, A)$ (respectively, $(K, L) \in ANE(X, A)$) provided any map $f : A \to L$ extends to $F : X \to K$ (respectively, extends to $F : U \to K$ for some neighborhood U of A in X). $K \in AE(X)$ (respectively, $K \in ANE(X)$) means that $(K, K) \in AE(X, A)$ (respectively, $(K, K) \in ANE(X, A)$) for all closed subsets A of X.

Which definitions/theorems of topology can be expressed using the concept of absolute (neighborhood) extensor? Here is a list:

7.18. PROPOSITION.

(1) *X is Hausdorff if and only if $S^0 \in ANE(X, A)$ for all finite subsets A of X.*
(2) *X is normal if and only if $S^0 \in ANE(X)$.*
(3) *X is collectionwise normal if and only if $D \in ANE(X)$ for all discrete spaces D.*
(4) *(Urysohn Lemma) If $S^0 \in ANE(X)$, then $(I, S^0) \in AE(X)$.*
(5) *(Tietze–Urysohn Theorem) $S^0 \in ANE(X)$ if and only if $I \in AE(X)$.*
(6) *A is C^*-embedded in X if and only if $I \in AE(X, A)$.*
(7) *A is C-embedded in X if and only if $R \in AE(X, A)$.*
(8) *A is P-embedded in X if and only if $M \in AE(X, A)$ for all complete AR's M.*
(9) *A is M-embedded in X if and only if $M \in AE(X, A)$ for all AR's M.*
(10) *$S^n \in AE(X)$ if and only if $\dim X \leqslant n$,*
(11) *$K(G, n) \in AE(X)$ if and only if $\dim_G X \leqslant n$.*
(12) *If $A \hookrightarrow X$ is a shape equivalence and A is closed in metrizable X, then $P \in AE(X, A)$ for all polyhedra P.*

(13) *If A is closed in the metrizable space X, then $A \hookrightarrow X$ is a strong shape equivalence if and only if $P \in AE(X, A)$ and $P \in AE(X \times I, A \times I \cup X \times \{0, 1\})$ for all polyhedra P.*

(14) *(X, A) has Homotopy Extension Property with respect to Y if and only if $Y \in AE(X \times I, A \times I \cup X \times \{0\})$.*

Can one generalize well-known theorems to the setting of extension theory? The first part of this section implies that joins of complexes are vital in the understanding of classical results of dimension theory. Therefore, it is not surprising that the simplest join of spaces, the cone (which is the join of a space with the one-point space) plays a significant role in general extension theory. The following is a generalization of the Tietze–Urysohn Extension Theorem for cones (previously known if Y is discrete).

7.19. THEOREM. *Suppose $Y \neq \{point\}$ is a Hausdorff space. Then, the following conditions are equivalent:*
(1) *$Cone(Y) \in AE(X)$.*
(2) *$(Cone(Y), Y) \in AE(X)$.*
(3) *$Y \in ANE(X)$.*

The following is a general version of the Homotopy Extension Theorem:

7.20. THEOREM. *Suppose A is a subspace of X and Y is a topological space. Then, the following conditions are equivalent:*
(1) *$Y \in AE(X \times I, X \times \{0\} \cup A \times I)$.*
(2) *$(Cone(Y), Y) \in AE(X \times I, X \times \{0\} \cup A \times I)$.*

To an algebraic/geometric topologist, maps into simplicial complexes (or CW complexes) are of importance. Natural problems are:

7.21. PROBLEM. Characterize spaces X such that $K \in ANE(X)$ for all simplicial complexes K.

7.22. PROBLEM. Characterize spaces X such that $K_{CW} \in ANE(X)$ for all simplicial complexes K. (Here K_{CW} is K equipped with the weak topology.)

It is well-known that each simplicial complex K has a natural partition of unity $\{\lambda_v\}_{v \in K^{(0)}}$, called barycentric coordinates. It is also known that $f : A \to K$ is continuous if and only if $\lambda_v \circ f$ is continuous for all $v \in K^{(0)}$. Obviously, in this case, $\{\lambda_v \circ f\}_{v \in K^{(0)}}$ is a point-finite partition of unity on A. This can be generalized as follows.

7.23. THEOREM. *A partition of unity $\{\alpha_s\}_{s \in S}$ on X determines a map $\alpha : X \to \mathrm{cl}(\Delta_S) \subset l^1_S$. A point-finite partition of unity $\{\alpha_s\}_{s \in S}$ on X determines a map $\alpha : X \to \Delta_S \subset l^1_S$. Here, Δ_S is the space of all points $(x_s)_{s \in S} \in l^1_S$ such that $x_s = 0$ for all but finitely many $s \in S$, $x_s \geqslant 0$ for all s, and $\sum_{s \in S} x_s = 1$.*

An solution to Problem 7.21 is:

7.24. THEOREM. *$K \in ANE(X, A)$ for all simplicial complexes K if and only if any point-finite partition of unity on A extends over X.*

An solution to Problem 7.22:

7.25. THEOREM. *Suppose X is first countable. $K_{CW} \in ANE(X, A)$ for all simplicial complexes K if and only if any locally-finite partition of unity on A extends over X.*

Naturally, one may wonder about spaces X such that any partition of unity on a closed subset A of X extends over X. It turns out that these are precisely the collectionwise normal spaces.

Let us demonstrate a few additional insights one gains by using partitions of unity. Suppose X is normal and has a σ-locally finite basis $\bigcup \mathcal{U}_n$. One can find a partition of unity α_n on X such that $Cozero(\alpha_n)$ contains $\mathcal{U}_n$. Now, $\alpha = \{\alpha_n/2^n\}_{n \geqslant 1}$ is a partition of unity on X such that $Cozero(\alpha)$ is a basis of X. By interpreting α as a map $\alpha : X \to \mathrm{cl}(\Delta_S)$ one easily gets that $\alpha : X \to \alpha(X)$ is a homeomorphism. In particular, X is metrizable. Thus, we get a new metrization criterion as follows.

7.26. THEOREM. *X is metrizable if and only if there is a partition of unity α on X such that $Cozero(\alpha)$ is a basis of X.*

Let Y be the convex hull of $\alpha(X)$ and suppose $y = c_1 \cdot \alpha(x_1) + c_2 \cdot \alpha(x_2)$ belongs to the closure of $\alpha(X)$ in Y. Choose $s, t \in S$ so that $x_1 \in Cozero(\alpha_s)$, $x_2 \in Cozero(\alpha_t)$, and $Cozero(\alpha_s) \cap Cozero(\alpha_t) = \emptyset$. Notice that $y \in Cozero(\lambda_s) \cap Cozero(\lambda_t)$ but $\alpha(X) \cap Cozero(\lambda_s) \cap Cozero(\lambda_t) = \emptyset$, a contradiction. Thus, one gets:

7.27. KURATOWSKI–WOJDYSŁAWSKI THEOREM. *A metrizable space embeds as a closed subset of a convex subset of a Banach space.*

By a similar argument one gets that all points in $\alpha(X)$ are linearly independent.

7.28. ARENS–EELLS THEOREM. *A metrizable space embeds as a closed subset of a normed linear space so that all points are linearly independent.*

Here is a result which unifies three well-known theorems:

7.29. THEOREM. *If X is metrizable, then any (locally-finite, point-finite) partition of unity on a closed subset of X extends over X.*

The case of arbitrary partitions of unity is closest to Dugundji's Extension Theorem (see its proof in [82, p. 35]).

7.30. DUGUNDJI'S THEOREM. *A convex subset of a normed vector space is an absolute extensor of all metric spaces.*

The case of point-finite partitions of unity corresponds to the fact that simplicial complexes are ANR's (see [82, p. 304]).

The case of locally-finite partitions of unity corresponds to Kodama's Theorem [70].

7.31. KODAMA'S THEOREM. *A CW complex is an absolute neighborhood extensor of all metric spaces.*

8. Cohomological dimension and its connections to other branches of topology/mathematics

In this section we review applications of cohomological dimension theory which should help the reader in assessing its place in modern topology.

8.1. *Compact group actions and Hilbert–Smith Conjecture*

Spaces with different cohomological dimensions for different Bockstein groups appear while considering free actions of $\mathbf{p}$-adic groups on n-manifolds.

8.1. DEFINITION OF THE $\mathbf{p}$-ADIC GROUP. Let $\mathbf{p}$ be a prime number. The $\mathbf{p}$-adic group $A_\mathbf{p}$ is the inverse limit of $\cdots \to \mathbf{Z}/\mathbf{p}^{n+1} \to \mathbf{Z}/\mathbf{p}^n \to \cdots \to \mathbf{Z}/\mathbf{p}$, where $\mathbf{Z}/\mathbf{p}^{n+1} \to \mathbf{Z}/\mathbf{p}^n$ is induced by the identity $\mathbf{Z} \to \mathbf{Z}$.

The main problem regarding actions of $\mathbf{p}$-adic groups is:

8.2. CONJECTURE. The $\mathbf{p}$-adic group $A_\mathbf{p}$ cannot act freely on an n-manifold.

Interest in Conjecture 8.2 is motivated by Smith's [102] generalization of Hilbert's Fifth Problem [66]. Hilbert asked whether a topological group G, the underlying space $|G|$ of which, is a (finite-dimensional) manifold, admits the structure of a Lie group, i.e., whether $|G|$ admits the structure of a real-analytic manifold on which group multiplication and inversion are analytic functions. Hilbert's problem was solved in the positive by von Neuman in 1933 [106] for compact groups and by Montgomery and Zippin in 1952 [87] for locally compact groups. Smith's generalization, which asks whether a compact topological group acting effectively (i.e., so that each element that is not the identity moves some point) on a finite-dimensional manifold must be a (subgroup of a) Lie group, is still unanswered. It is now known as the Hilbert–Smith Conjecture. The Hilbert–Smith Conjecture is known to be equivalent to the conjecture that a compact topological group G acting effectively on a compact finite-dimensional manifold M^n cannot have arbitrarily small subgroups; moreover, one may restrict attention to the case that G is totally disconnected, so it is equivalent to the proposition that if G is zero-dimensional, then it is finite (see Section 10 of [102]). Work of Newman [86] and Smith [101] eliminates arbitrarily small torsion and reduces the problem to consideration of the p-adic group (see Section 10 of [102]). Thus, Hilbert–Smith Conjecture is equivalent to Conjecture 8.2.

So far the best result regarding the Hilbert–Smith Conjecture is due to Yang [113].

8.3. YANG THEOREM. Suppose $\mathbf{p}$ is a prime and the $\mathbf{p}$-adic group $A_{\mathbf{p}}$ acts freely on an n-manifold M^n. Then,

(1) $\dim_{\mathbf{Z}}(M^n/A_{\mathbf{p}}) = n+2$,
(2) $\dim_{\mathbf{Z}/\mathbf{p}}(M^n/A_{\mathbf{p}}) = n+1$,
(3) $\dim_{\mathbf{Z}/\mathbf{q}}(M^n/A_{\mathbf{p}}) = n$ if $\mathbf{q} \neq \mathbf{p}$ is a prime, and
(4) $\dim_{\mathbf{Q}}(M^n/A_{\mathbf{p}}) = n$.

In the case of a locally compact, Hausdorff, homologically n-dimensional space X supporting an effective A_p action Yang proved that $\dim_{\mathbf{Z}}(X/A_p) \leqslant n+3$. Actions of this kind are called *dimension-raising*. Examples of interesting dimension-raising actions of A_p on a non-manifold X with $\dim(X/G) = \dim(X)+1$ or $\dim(X/G) = \dim(X)+2$ were constructed in [73] and [94] (see [111,112]), and an example of a free A_p-action on a 2-dimensional cell-like set was constructed by Bestvina and Edwards. Dimension-raising actions of torsion groups on compact metric spaces were recently constructed by Dranishnikov and West [44].

8.4. THEOREM ([44]). *For each integer $n \geqslant 3$ and each prime $\mathbf{p}$ there is an action of $G = \prod_{i=1}^{\infty}(Z/\mathbf{p})_i$ on a compact, 2-dimensional metric space X such that* $\dim(X/G) = n$; *moreover,* $\dim(X \times X) = 3$.

8.5. THEOREM ([44]). *For each compact metric space Y, and each prime $\mathbf{p}$, Y is the orbit space of an action of $G = \prod_{i=1}^{\infty}(Z/\mathbf{p})_i$ on a metric compactum X with* $\dim_{\mathbf{Z}/\mathbf{p}}(X) = 1$.

8.2. *Gromov–Hausdorff metric and integral dimension*

Cell-like maps and spaces of finite integral cohomological dimension show up when discussing limits of manifolds in the Gromov–Hausdorff metric.

8.6. DEFINITION OF HAUSDORFF DISTANCE. Let (Z,d) be a compact metric space. Given a subset A of Z its ε-neighborhood is denoted by $N_\varepsilon(A)$. If A, B are closed subsets of Z, then the *Hausdorff distance* $d_Z^H(A,B)$ from A to B is defined as

$$d_Z^H(A,B) = \inf\{\varepsilon > 0 \mid A \subset N_\varepsilon(B),\ B \subset N_\varepsilon(A)\}.$$

8.7. DEFINITION OF GROMOV–HAUSDORFF DISTANCE. If A, B are compact metric spaces, then the *Gromov–Hausdorff distance* $d_{GH}^H(A,B)$ from A to B is defined as

$$d_{GH}(A,B) = \inf\{d_Z^H(A,B) \mid Z \text{ contains isometric copies of } A, B\}.$$

One can consider the set $\mathcal{CM}$ of isometry classes of compact metric spaces with the Gromov–Hausdorff metric. It turns out that $\mathcal{CM}$ is a complete space (see [64] or [91] for an exposition). Within $\mathcal{CM}$ one can consider a subset $\mathcal{M}(n,\rho)$ consisting of all compacta of dimension at most n and whose contractibility function is ρ.

8.8. DEFINITION OF A CONTRACTIBILITY FUNCTION. Suppose X is a compact metric space of dimension at most n and $\rho : [0, R] \to [0, \infty)$ is a function such that $\rho(0) = 0$, ρ is continuous at 0, and $\rho(t) \geqslant t$ for all $t \leqslant R$. We say that ρ is a *contractibility function* of X provided any map $\alpha : S^k \to X$ with $k \leqslant n$ and $\operatorname{diam}(\alpha) < \varepsilon \leqslant R$ extends to $\beta : B^{k+1} \to X$ with $\operatorname{diam}(\beta) < \rho(\varepsilon)$.

It turns out that the closure of $\mathcal{M}(n, \rho)$ is contained in the subset consisting of compacta of integral cohomological dimension at most n.

8.9. THEOREM ([84]). *If $X \in \operatorname{cl}(\mathcal{M}(n, \rho))$, then X is a cell-like image of $Y \in \mathcal{M}(n, \rho)$.*

There is a partial converse to the above theorem.

8.10. THEOREM ([84]). *Suppose X is a cell-like image of a compact n-manifold M^n. There is a contractibility function ρ and a path $\omega : [0, 1] \to \operatorname{cl}(\mathcal{M}(n, \rho))$ such that $\rho(1) = X$ and $\rho(t)$ is homeomorphic to M^n for $t < 1$.*

8.3. *Cohomological dimension of groups*

In this section we will discuss the relation between the cohomological dimension of a CAT(0) group and the cohomological dimension of its boundary. For a more detailed discussion see [6,37], and [35].

8.11. DEFINITION OF COHOMOLOGICAL DIMENSION OF A GROUP. The *cohomological dimension* $\operatorname{cd}_R(\Gamma)$ of a group Γ with respect to a ring R is

$$\sup\{n \mid H^n(\Gamma; M) \neq 0 \text{ for some } R\Gamma\text{-module } M\}.$$

Here $H^n(\Gamma; M)$ is defined as $H^n(K(\Gamma, 1); \mathcal{M})$, where $\mathcal{M}$ is the local coefficient system generated by the action of Γ on M.

If $R = \mathbf{Z}$ is the ring of integers, then $\operatorname{cd}_R(\Gamma)$ is denoted by $\operatorname{cd}(\Gamma)$ and is called the cohomological dimension of the group Γ.

The *geometric dimension* $\dim(\Gamma)$ of Γ is the minimal dimension of CW complexes representing $K(\Gamma, 1)$.

If Γ is finitely presented and of finite geometric dimension, then $\operatorname{cd}_R(\Gamma)$ coincides with $\max\{n \mid H^n(\Gamma; R\Gamma) \neq 0\} = \max\{n \mid H^n_c(X; R) \neq 0\}$, where X is the universal covering space of $K(\Gamma, 1)$, and $H^n_c(X; R)$ is the cohomology group with compact supports, which coincides with the Čech cohomology group of the one-point compactification of X.

8.12. DEFINITION OF VIRTUAL COHOMOLOGICAL DIMENSION OF A GROUP. The *virtual cohomological dimension* $\operatorname{vcd}_R(\Gamma)$ of a group Γ with respect to a ring R is

$$\min\{\operatorname{cd}_R(\Gamma') \mid \Gamma' \text{ is a subgroup of } \Gamma \text{ of finite index}\}.$$

8.13. Definition of a Coxeter group. A *Coxeter group* is a group with presentation

$$\langle v \mid v^2 = 1 \text{ for all } v \in V,\ (vw)^{m(v,w)} = 1 \text{ for all } (v,w) \in I \rangle$$

where V is a finite set, I is a symmetric subset of $V \times V$, and $m: I \to \mathbf{Z}_+$ is a symmetric function into the set of positive integers. (Γ, V) is called a *Coxeter system*.

If $m \equiv 2$, then Γ is called a *right-angled Coxeter group*.

It is known that Coxeter groups have finite virtual cohomological dimension.

8.14. Definition of CAT(0) geodesic space. Let (X, d) be a metric space such that for every two points $x, y \in X$ there is an isometry $g: [0, d(x, y)] \to X$ (g is called a *geodesic*) so that $g(0) = x$ and $g(d(x, y)) = y$. Given three points $x_i \in X$, $i = 1, 2, 3$, choose an isometry $i: \{x_1, x_2, x_3\} \to R^2$. If, for every y lying on a geodesic joining x_2 and x_3, $d(x_1, y) \leqslant |i(x_1) - y'|$, where y' is the corresponding point on the segment joining $i(x_2)$ and $i(x_3)$, then X is called a CAT(0) *space*.

8.15. Definition of CAT(0) group. A group Γ is called a CAT(0) *group* if the universal cover of a $K(\Gamma, 1)$ admits a CAT(0) metric so that Γ acts on X via isometries.

8.16. Definition of the boundary of a CAT(0) space. Given a CAT(0) space (X, d) and given a point $x \in X$, one considers all isometric embeddings $g: [0, \infty) \to X$ (called *geodesic rays*) so that $g(0) = x$. The *visual sphere* $S_x(\infty)$ at x is the set of all geodesic rays emanating from x with the topology of uniform convergence on all compact sets. It turns out that the topological type of all visual spheres is the same regardless of x, and this is called the *boundary* of (X, d).

8.17. Definition of a boundary of a CAT(0) group. A *boundary* of a CAT(0) group Γ is the boundary of a CAT(0) universal cover of $K(\Gamma, 1)$.

In the case of a right-angled Coxeter group Γ there is a specific construction of the universal cover of a $K(\Gamma, 1)$ which admits a CAT(0) metric. Thus, in this case, the boundary of that specific universal cover is called the boundary of Γ. Now we arrive at the connection between the cohomological dimension of groups and the cohomological dimension of compacta.

8.18. Bestvina–Mess Theorem. *Suppose Γ is a finitely presented* CAT(0) *group of finite geometric dimension, and let R be a ring with unity. Then,* $\dim_R(\partial\Gamma) = \mathrm{cd}_R(\Gamma) - 1$.

This result was used by Dranishnikov to construct interesting examples of boundaries of CAT(0) groups.

8.19. Theorem. *Let $\mathcal{F}$ be the family of additive groups $\{\mathbf{Q}\} \cup \bigcup_{\mathbf{p}\ prime}\{\mathbf{Z}/\mathbf{p}\}$. For every function $d: \mathcal{F} \to \{2, 3\}$ such that $d(\mathbf{Q}) = 2$ and $d(\mathbf{Z}/\mathbf{p}) = 2$ for all but finitely many $\mathbf{p}$, there is a Coxeter group Γ with* $\mathrm{vcd}(\Gamma) = 3$ *and* $d(G) = \mathrm{vcd}_G(\Gamma)$ *for all $G \in \mathcal{F}$.*

8.20. COROLLARY. *For every prime* $\mathbf{p}$ *there is a* CAT(0) *group* $\Gamma_{\mathbf{p}}$ *whose boundary* $\partial\Gamma_{\mathbf{p}}$ *is* 2*-dimensional,* $\dim_{\mathbf{Z}/\mathbf{p}}(\partial\Gamma_{\mathbf{p}}) = 2$, $\dim_{\mathbf{Q}}(\partial\Gamma_{\mathbf{p}}) = 1$, *and* $\dim_{\mathbf{Z}/\mathbf{q}}(\partial\Gamma_{\mathbf{p}}) = 1$ *for every prime* $\mathbf{q} \neq \mathbf{p}$.

Further information on CAT(0) spaces and CAT(0) groups may be found in the chapters of this book by Cannon [17] and Davis [22].

8.4. *Mapping intersection problem*

Recently there was much research done on the so-called Mapping Intersection Problem (see Conjecture 8.22). We recommend [34] as a source for the latest results.

8.21. DEFINITION OF STABLE/UNSTABLE INTERSECTION OF MAPS. Suppose $f: X \to R^n$ and $g: Y \to R^n$ are two maps of compacta into Euclidean n-space. If there exists $\varepsilon > 0$ such that $f'(X) \cap g'(Y) \neq \emptyset$ for all $f': X \to R^n$, $g': Y \to R^n$ satisfying $|f - f'| < \varepsilon$, $|g - g'| < \varepsilon$, then f and g are said to have *stable intersection*. If such ε does not exist, then f and g are said to have *unstable intersection*.

8.22. CONJECTURE. Suppose X, Y are compacta. There is a pair of maps $f: X \to R^n$ and $g: Y \to R^n$ having stable intersection if and only if $\dim(X \times Y) \geqslant n$.

8.23. DRANISHNIKOV'S THEOREM. *Suppose* $f: X \to R^n$ *and* $g: Y \to R^n$ *is a pair of maps of compacta such that* $\dim(X) < n - 2$ *and* $\dim(Y) \leqslant n - 2$. *For any* $\varepsilon > 0$ *there exist maps* $f': X \to R^n$, $g': Y \to R^n$ *satisfying* $|f - f'| < \varepsilon$, $|g - g'| < \varepsilon$, *and* $\dim(f'(X) \cap \dim g'(Y)) \leqslant \dim(X \times Y) - n$.

Theorem 8.23 represents the latest achievement in a series of results aiming at solving Conjecture 8.22. The main ingredient distinguishing it from the preceding research is that it applied Theorem 7.8.

8.5. *Rings of complex and real functions on compact metric spaces*

Given a compactum X one can consider algebraic properties associated with the ring $C_{\mathbf{R}}(X)$ (respectively, $C_{\mathbf{C}}(X)$) of all real-valued (respectively, complex-valued) continuous functions on X. It turns out that certain properties of rings $M_n(C_{\mathbf{R}}(X))$ of $n \times n$-matrices over $C_{\mathbf{R}}(X)$ can be expressed in terms of the rational dimension of X.

8.24. DEFINITION OF POWER SUBSTITUTION PROPERTY. A ring R with unity has the *power substitution property* provided $a \cdot x + b \cdot y = 1$ implies the existence of n so that $a \cdot 1_n + b \cdot M_n(R)$ contains a unit. If n is the same for all choices of a, b, then we say that R has the *n-power substitution property*.

8.25. DEFINITION OF WEAK POWER SUBSTITUTION PROPERTY. A ring R with unity has the *weak power substitution property* provided $a \cdot R + I = R$ for some ideal I of R implies the existence of n so that $a \cdot 1_n + b \cdot M_n(I)$ contains a unit.

For general rings, the power substitution property implies the weak power substitution property. In the case of commutative rings, they are equivalent.

8.26. THEOREM ([33]). *Let X be a compact metric space. The following conditions are equivalent*:
(1) *$M_n(C_{\mathbf{R}}(X))$ has the weak power substitution property for some $n \geqslant 3$.*
(2) $\dim_{\mathbf{Q}}(X) \leqslant 3$.
(3) *$M_n(C_{\mathbf{R}}(X))$ has the weak power substitution property for all n.*

8.27. THEOREM ([33]). *Let X be a compact metric space. The following conditions are equivalent*:
(1) *$M_n(C_{\mathbf{C}}(X))$ has the weak power substitution property for some n.*
(2) $\dim_{\mathbf{Q}}(X) \leqslant 1$.
(3) *$M_n(C_{\mathbf{C}}(X))$ has the weak power substitution property for all n.*

8.28. THEOREM ([33]). *For every n and every prime* $\mathbf{p}$, *there is an n-dimensional compactum X so that $C_{\mathbf{C}}(X)$ has the $\mathbf{p}^{n-1}$-power substitution property.*

8.6. *Finitistic spaces*

The classical cohomological methods in the study of group actions were applied either to compact Hausdorff spaces or paracompact spaces of finite cohomological dimension (see [9] and [13]). Swan [105] introduced the concept of finitistic spaces and obtained results generalizing classical Smith-type fixed point theorems.

8.29. DEFINITION OF FINITISTIC SPACES. A space X is *finitistic* if every open cover of X has an open refinement of finite order.

Recall that the order of an open cover is at most n if every $n+2$ distinct elements of the cover have empty intersection. Since finite open covers have finite order, compact Hausdorff spaces are finitistic. Also, paracompact spaces of finite covering dimension are finitistic. Indeed, $\dim(X) \leqslant n$, where $0 \leqslant n < \infty$, if every finite open cover of X has a finite open refinement of order at most n, and in the case of a paracompact space X one can find open an refinement of order at most n for every open cover of X (see [63, Lemma 3.1.9 on p. 172]). Thus, the class of paracompact, finitistic spaces may be formally viewed as a natural class combining both compact and finite-dimensional spaces.

After Swan's introduction of finitistic spaces, Bredon [13] set the trend of stating results on the cohomological structure of fixed point sets in terms of finitistic spaces, and it is apparent that finitistic spaces give a natural environment for using the Čech method in generalizing cohomological results on fixed point sets and orbit spaces. As seen in [24], a typical result on toral actions and rational coefficients involves the assumption that both the total space X and the orbit space X/T^n (T^n being the n-torus) are finitistic. In an effort to weaken those assumptions, Deo and Tripathi proved the following.

8.30. THEOREM ([25]). *Suppose a compact Lie group G acts on a paracompact, finitistic space X. Then, the orbit space X/G is finitistic.*

The converse of that statement was proved by Deo and Singh.

8.31. THEOREM ([23]). *Suppose a compact Lie group G acts on a paracompact X. If the orbit space X/G is finitistic, then X is finitistic, too.*

Another branch of topology where finitistic spaces surfaced recently is the cohomological dimension theory. Namely, Rubin and Schapiro obtained the following.

8.32. THEOREM ([97]). *Suppose X is a paracompact, finitistic space and G is a finitely generated, abelian group. Then*

(a) $\dim_G(X) = \dim_G(\beta X)$, *where βX is the Čech–Stone compactification of X, and*

(b) *if X is metric, separable, finitistic space, then X has a metric compactification preserving cohomological dimension.*

Earlier, Dranishnikov [27] found a metric, separable space X so that $\dim_{\mathbf{Z}}(X) = 4$ and $\dim_{\mathbf{Z}}(\beta X) > 4$, and Dydak and Walsh [56] found a metric, separable space X so that $\dim_{\mathbf{Z}}(X) = 4$ and $\dim_{\mathbf{Z}}(\kappa X) > 4$ for any compactification κX of X.

In a recent paper (see [52]) Dydak, Mishra, and Shukla generalized the above mentioned results of [25,23], and [97]. This was achieved by applying the methodology of cohomological dimension theory in the form of approximate liftings (see Definition 3.2). One can specialize approximate liftings to inclusion of skeleta as follows.

8.33. DEFINITION OF K-APPROXIMATIONS. Let K be a metric simplicial complex. A map $g : X \to K$ is a *K-approximation* of $f : X \to K$ provided for each simplex Δ of K and each $x \in X$, $f(x) \in \Delta$ implies $g(x) \in \Delta$. g is an *n-dimensional* (respectively, *finite-dimensional*) *K-approximation* of f if it is a K-approximation and $g(X) \subset K^{(n)}$ (respectively, $g(X) \subset K^{(m)}$ for some m).

Using K-approximations one obtains a useful characterization of finitistic spaces.

8.34. THEOREM ([52]). *For a paracompact space X the following conditions are equivalent:*

(a) *X is finitistic.*

(b) *For any metric simplicial complex K every map $f : X \to K$ has a finite-dimensional K-approximation g.*

(c) *For any metric simplicial complex K and every $m \geqslant -1$, every map $f : X \to K$ has a finite-dimensional K-approximation g so that $g|f^{-1}(K^{(m)}) = f|f^{-1}(K^{(m)})$.*

The above characterization can be used to prove another interesting characterization of finitistic spaces.

8.35. THEOREM ([65] and [52]). *A paracompact space X is finitistic if and only if there is a compact subset Z of X so that $X - U$ is finite-dimensional for every open neighborhood U of Z in X.*

8.36. COROLLARY. *Suppose A is finitistic and a closed subset of a paracompact space X. If $X - U$ is finitistic for every open neighborhood U of A in X, then X is finitistic.*

8.37. THEOREM. *Suppose X is a separable metric space and A is a finitistic subset of X. There is a G_δ-subset B of X containing A so that B is finitistic.*

In this section we generalize some results of classical dimension theory on maps which raise or lower covering dimension. The following two fundamental theorems can be found in [63, pp. 196–200].

8.38. THEOREM ON DIMENSION-RAISING MAPPINGS FOR DIM. *Suppose $f : X \to Y$ is a closed, surjective map of normal spaces such that there is an integer k with $f^{-1}(y)$ containing at most k elements for all $y \in Y$. Then,* $\dim(Y) \leqslant \dim(X) + k - 1$.

8.39. THEOREM ON DIMENSION-LOWERING MAPPINGS FOR DIM. *Suppose $f : X \to Y$ is a closed, surjective map of a normal space X onto a weakly paracompact normal space Y such that there is an integer k with* $\dim(f^{-1}(y)) \leqslant k$ *for all $y \in Y$. Then,* $\dim(X) \leqslant \dim(Y) + k$.

First, we generalize the well-known fact that if $f : X \to Y$ is perfect and Y is compact, then X is compact:

8.40. THEOREM. *Suppose $f : X \to Y$ is a closed map of paracompact spaces such that $f^{-1}(y)$ is finitistic for all $y \in Y$. If Y is compact, then X is finitistic.*

One cannot weaken the assumptions of Theorem 8.40 by assuming that Y is finitistic rather than compact. Indeed, let Q be the Hilbert cube. The projection $f : Q \times \mathbf{Z} \to \mathbf{Z}$ has compact fibers, $\mathbf{Z}$ being discrete is finitistic, and $Q \times \mathbf{Z}$ is not finitistic (use Theorems 2.4 or 0.6). However, we may put an additional restriction on the fibers of f as seen in the next result.

8.41. THEOREM. *Suppose $f : X \to Y$ is a closed map of paracompact spaces such that $f^{-1}(A)$ is finite-dimensional for all finite-dimensional closed subsets A of Y. If B is a closed, finitistic subset of Y, then $f^{-1}(B)$ is finitistic.*

Here is a generalization of Theorem 8.39.

8.42. COROLLARY. *Suppose $f : X \to Y$ is a closed map of paracompact spaces such that there is an integer $k \geqslant 0$ with* $\dim(f^{-1}(y)) \leqslant k$ *for all $y \in Y$. If Y is finitistic, then so is X.*

As an application we get a generalization of a result of Deo and Singh [23] (see Theorem 0.4) regarding compact Lie group actions. The generalization applies to all compact group actions.

8.43. COROLLARY. *Suppose a compact, finite-dimensional group G acts on a paracompact space X. If X/G is finitistic, then so is X.*

Notice that the converse implication (i.e., if X is finitistic, then X/G is finitistic), known to be true for Lie groups G (see [25] and Theorem 8.30), does not hold for general compact and finite-dimensional groups G. A counterexample to that implication can be easily constructed using Theorem 8.4 of Dranishnikov and West. Indeed, $X = \bigoplus_{n=1}^{\infty} X_n$ (the discrete sum of all X_n) is 2-dimensional and admits an action of G so that X/G is not finitistic.

Let us show how to reduce the result of Deo and Tripathi (see Theorem 8.30) to the finite-dimensional case.

8.44. THEOREM. *Suppose G is a compact topological group. The following conditions are equivalent*:

(a) *For any action of G on a finite-dimensional, paracompact space X, the orbit space X/G is finite-dimensional.*

(b) *For any action of G on a finitistic, paracompact space X, the orbit space X/G is finitistic.*

Since compact Lie groups are known to satisfy condition (a) of Theorem 8.44, Theorem 8.30 of Deo and Tripathi [25] follows.

8.45. THEOREM. *Suppose $f : X \to Y$ is a closed map of paracompact spaces such that $f(A)$ is finite-dimensional for all finite-dimensional closed subsets A of X. If B is a closed, finitistic subset of X, then $f(B)$ is finitistic.*

The following is a generalization of Theorem 8.38.

8.46. COROLLARY. *Suppose $f : X \to Y$ is a closed, surjective map of paracompact spaces such that there is an integer k with $f^{-1}(y)$ containing at most k elements for all $y \in Y$. If X is finitistic, then so is Y.*

The following is a generalization of Theorem 8.32.

8.47. THEOREM. *Suppose X is a paracompact, finitistic space and K is a metric simplicial complex which is homotopy equivalent to a CW complex L whose skeleta are finite. If K is an absolute extensor of X, then K is an absolute extensor of the Čech–Stone compactification βX of X. If K is an absolute extensor of βX, and is complete, then K is an absolute extensor of X.*

9. Open problems

This section is devoted to some open problems in extension theory. There is one case of the Cell-like Mapping Problem which remains unsolved in spite of considerable effort.

9.1. PROBLEM. Can cell-like maps on closed 4-manifolds raise dimension?

It is known that ANR's of finite integral cohomological dimension are finite-dimensional. Therefore, specializing Problem 0.8 to ANR's makes sense only in case of $G \neq \mathbf{Z}$. It turns out that the case $G = \mathbf{Q}$ is the most important one.

9.2. PROBLEM ([27]). Is there an ANR of infinite dimension but of finite rational cohomological dimension?

More generally, one may seek analogs of the Dranishnikov Realization Theorem for ANR's.

9.3. PROBLEM ([27]). Characterize dimension functions of compact (respectively, separable) ANR's.

In the case of extension dimension/types the following four problems seem relevant.

9.4. PROBLEM ([38]). Characterize CW complexes whose extension universe has a compact metric representative.

9.5. PROBLEM ([38]). Characterize compacta whose extension dimension has a representative being a countable CW complex.

9.6. PROBLEM ([38]). Characterize CW complexes whose extension universe has a representative being a compact ANR.

9.7. PROBLEM ([38]). Let $\mathcal{M}$ be the class of all countable CW complexes.

(1) Given $X \in \mathcal{SEPARABLE}$, is there a minimum in

$$\{[M] \in \mathrm{ExtTypes}(\mathcal{SEPARABLE}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}?$$

(2) Given $X \in \mathcal{COMPACTA}$, is there a minimum in

$$\{[M] \in \mathrm{ExtTypes}(\mathcal{COMPACTA}, \mathcal{M}) \mid \text{ext-dim}(X) \leqslant M\}?$$

Conjecture 7.7 remains unsolved. Here are the most important special cases of 7.7:

9.8. PROBLEM ([109]). Let $n > 1$. Is there a universal space in the class of compacta of integral cohomological dimension n?

9.9. PROBLEM. Let $n > 0$. Is there a universal space in the class of compacta of rational cohomological dimension n?

9.10. PROBLEM. Is there a universal space in the class of compacta of rational cohomological dimension 1 and covering dimension 2?

It is known that, in general, one cannot compactify a metric, separable space of integral dimension 4 without increasing the integral cohomological dimension. It would be of interest to resolve the following question:

9.11. PROBLEM ([56]). Suppose X is a metric, separable space of finite integral dimension. Is there a compactum Y containing X so that the integral dimension of Y is finite?

The next two problems deal with the cohomological dimension of Coxeter groups:

9.12. PROBLEM ([37]). Characterize dimension functions d which can be realized by Coxeter groups. Is there a Coxeter group Γ_n so that $\mathrm{vcd}_{\mathbf{Q}}(\Gamma_n) = 2$ and $\mathrm{vcd}(\Gamma_n) = n$ for each large n?

9.13. PROBLEM ([37]). Suppose $n \geqslant 3$. Is there a Coxeter group Γ_n so that $\dim_{\mathbf{Q}}(\partial\Gamma_n) = 1$ and $\dim(\partial\Gamma_n) = n$?

References

[1] P.S. Alexandroff, *Dimensionstheorie, ein Beitrag zur Geometrie der abgeschlossenen Mengen*, Math. Ann. **106** (1932), 161–238 (in German).

[2] P.S. Alexandroff, *Einige Problemstellungen in der mengentheoretischen Topologie*, Mat. Sbor. **43** (1936), 619–634 (in German).

[3] P.S. Alexandroff, *Introduction to Homological Dimension Theory and General Combinatorial Topology*, Nauka, Moscow (1975) (in Russian).

[4] D.W. Anderson and L. Hodgkin, *The K-theory of Eilenberg–MacLane complexes*, Topology **7** (1968), 317–329.

[5] V.M. Buchstaber and A.S. Mischenko, *K-theory on the category of infinite cell complexes*, Math. USSR Izv. **32** (2) (1968), 560–604.

[6] M. Bestvina and G. Mess, *The boundary of negatively curved groups*, J. Amer. Math. Soc. **4** (1991), 469–481.

[7] B.F. Bockstein, *Homological invariants of topological spaces I*, Trudy Moskov. Mat. Obshch. **5** (1956), 3–80 (in Russian; English translation in Amer. Math. Soc. Transl. **11** (3) (1050)).

[8] B. Booss and D.D. Bleecker, *Topology and Analysis*, Springer, Berlin (1985).

[9] A. Borel et al., *Seminar on Transformation Groups*, Ann. Math. Studies, Vol. 46, Princeton University Press, NJ (1960).

[10] V.G. Boltyanskij, *An example of two-dimensional compactum whose topological square is three-dimensional*, Dokl. Akad. Nauk SSSR **67** (1949), 597–599.

[11] V.G. Boltyanskij, *On dimensional full-valuedness of compacta*, Dokl. Akad. Nauk SSSR **67** (1949), 773–777.

[12] K. Borsuk, *Un theoreme sur la prolongements des transformations*, Fund. Math. **29** (1937), 161–166.

[13] G.E. Bredon, *Introduction to compact transformation groups*, Academic Press, New York (1972).

[14] G.E. Bredon, *Sheaf Theory* (2nd edition), Springer, New York (1997).

[15] G.E. Bredon, F. Raymond and R.F. Williams, *p-Adic groups of transformations*, Trans. Amer. Math. Soc. **99** (1961), 488–498.
[16] L.E.J. Brouwer, *Über den naturlichen Dimensionsbegriff*, J. Reine Angew. Math. **142** (1913), 146–152 (in German).
[17] J.W. Cannon, *Geometric group theory*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 261–305.
[18] A. Chigogidze, *Cohomological dimension of Tychonoff spaces*, Topology Appl. **79** (1997), 197–228.
[19] A. Chigogidze, *Infinite-dimensional topology and shape theory*, Handbook of Geometric Topology, R.J. Davermann and R.B. Sher, eds, Elsevier, Amsterdam (2002), 307–371.
[20] H. Cohen, *A cohomological definition of dimension for locally compact Hausdorff spaces*, Duke Math. J. **21** (1954), 209–224.
[21] R.J. Daverman, *Decompositions of Manifolds*, Academic Press (1986).
[22] M.W. Davis, *Nonpositive curvature and reflection groups*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 373–422.
[23] S. Deo and T.B. Singh, *On the converse of some theorems about orbit spaces*, J. London Math. Soc. **25** (1982), 162–170.
[24] S. Deo, T.B. Singh and R.A. Shukla, *On an extension of localization theorem and generalized Conner conjecture*, Trans. Amer. Math. Soc. **269** (1982), 395–402.
[25] S. Deo and H.S. Tripathi, *Compact Lie group actions on finitistic spaces*, Topology **21** (1982), 393–399.
[26] A. Dold and R. Thom, *Quasifaserungen und unendliche symmetrische Produkte*, Ann. of Math. **67** (1958), 239–281 (in German).
[27] A.N. Dranishnikov, *Homological dimension theory*, Russian Math. Surveys **43** (4) (1988), 11–63.
[28] A.N. Dranishnikov, *On a problem of P.S. Alexandroff*, Math. USSR Sbor. **63** (2) (1988) 412–426, Mat. Sbor. **135** (4) (1988), 551–557.
[29] A.N. Dranishnikov, *Cohomological dimension is not preserved by Stone–Čech compactification*, C. R. Bulgarian Acad. Sci. **41** (1988), 9–10 (in Russian).
[30] A.N. Dranishnikov, *On intersection of compacta in euclidean space II*, Proc. Amer. Math. Soc. **113** (1991), 1149–1154.
[31] A.N. Dranishnikov, *Extension of maps into CW complexes*, Mat. Sborn. **182** (9) (1991), 1300–1310 (Math. USSR Sbor. **74** (1993), 47–56).
[32] A.N. Dranishnikov, *Eilenberg–Borsuk theorem for maps into arbitrary complexes*, Mat. Sbor. **185** (1994), 81–90.
[33] A.N. Dranishnikov, *The power substitution for rings of complex and real functions on compact metric spaces*, Proc. Amer. Math. Soc. **123** (1995), 2887–2893.
[34] A.N. Dranishnikov, *On the mapping intersection problem*, Pacific J. Math. **173** (2) (1996), 403–412.
[35] A.N. Dranishnikov, *On the virtual cohomological dimensions of Coxeter groups*, Proc. Amer. Math. Soc. **125** (1997), 1885–1891.
[36] A.N. Dranishnikov, *On dimension of the product of two compacta and the dimension of their intersection in general position in Euclidean space*, Trans. Amer. Math. Soc. **352** (2000), 5599–5618.
[37] A.N. Dranishnikov, *On boundaries of hyperbolic Coxeter groups*, Topology Appl. **111** (2001), 343–353.
[38] A. Dranishnikov and J. Dydak, *Extension dimension and extension types*, Proc. Steklov Institute of Mathematics, Vol. 212 (1996), 55–88.
[39] A. Dranishnikov and J. Dydak, *Extension theory of separable metrizable spaces with applications to dimension theory*, Trans. Amer. Math. Soc. **353** (2001), 133–156.
[40] A. Dranishnikov, J. Dydak and J. Walsh, *Cohomological dimension theory with applications*, Unpublished notes (1992).
[41] A. Dranishnikov and D. Repovš, *The Urysohn–Menger Sum Formula: An extension of the Dydak–Walsh theorem to dimension one*, J. Austral. Math. Soc., Ser. A **59** (1995), 273–282.
[42] A.N. Dranishnikov, D. Repovš and E.Ščepin, *Dimension of products with continua*, Topology Proc. **18** (1993), 57–73.
[43] A. Dranishnikov, D. Repovš and E. Ščepin, *On the failure of the Urysohn–Menger sum formula for cohomological dimension*, Proc. Amer. Math. Soc. **120** (1994), 1267–1270.
[44] A. Dranishnikov and J. West, *Compact group action that raise the dimension to infinity*, Topology Appl. **80** (1997), 101–114.

[45] J. Dydak, *Cohomological dimension and metrizable spaces*, Trans. Amer. Math. Soc. **337** (1993), 219–234.
[46] J. Dydak, *Union theorem for cohomological dimension: A simple counterexample*, Proc. Amer. Math. Soc. **121** (1994), 295–297.
[47] J. Dydak, *Cohomological dimension and metrizable spaces, II*, Trans. Amer. Math. Soc. **348** (1996), 1647–1661.
[48] J. Dydak, (in preparation) (1996).
[49] J. Dydak, *Extension theory: The interface between set-theoretic and algebraic topology*, Topology Appl. **20** (1996), 1–34.
[50] J. Dydak, *Covariant and contravariant approaches in topology*, Internat. J. Math. and Mathematical Sci. **20** (1997), 621–626.
[51] J. Dydak, *Covariant and contravariant points of view in topology with applications to function spaces*, Topology Appl. **94** (1999), 87–125.
[52] J. Dydak, S.N. Mishra and R.A. Shukla, *On finitistic spaces*, Topology Appl. (to appear).
[53] J. Dydak and J. Mogilski, *Universal cell-like maps*, Proc. Amer. Math. Soc. **122** (1994), 943–948.
[54] E. Dyer, *On dimension of products*, Fund. Math. **47** (1959), 141–160.
[55] J. Dydak and J. Segal, *Shape Theory: An Introduction*, Lecture Notes in Math., Vol. 688, Springer, Berlin (1978).
[56] J. Dydak and J.J. Walsh, *Spaces without cohomological dimension preserving compactifications*, Proc. Amer. Math. Soc. **113** (1991), 1155–1162.
[57] J. Dydak and J.J. Walsh, *Aspects of cohomological dimension for principal ideal domains*, Preprint.
[58] J. Dydak and J.J. Walsh, *Dimension theory and the Sullivan conjecture*, Topology Hawaii, K.H. Dovermann, ed., World Scientific, Singapore (1992), 75–89.
[59] J. Dydak and J.J. Walsh, *Infinite-dimensional compacta having cohomological dimension two: An application of the Sullivan Conjecture*, Topology **32** (1993), 93–104.
[60] J. Dydak and J.J. Walsh, *Complexes that arise in cohomological dimension theory: A unified approach*, J. London Math. Soc. **48** (1993), 329–347.
[61] S. Eilenberg, *Un theoreme de la dualite*, Fund. Math. **26** (1936), 280–282.
[62] R. Engelking, *General Topology*, Berlin (1989).
[63] R. Engelking, *Theory of Dimensions Finite and Infinite*, Sigma Series in Pure Math., Vol. 10, Heldermann (1995).
[64] K. Grove, *Metric differential geometry*, Differential Geometry, Proc. Nordic Summer School, Lyngby 1985, V.L. Hansen, ed., Springer Lecture Notes in Math., Vol. 1263, Springer, Berlin (1987), 171–227.
[65] Y. Hattori, *A note on finitistic spaces*, Questions Answers Gen. Topology **3** (1985), 47–55.
[66] D. Hilbert, *Mathematische Probleme*, Nachr. Akad. Wiss. Göttingen (1900), 253–297.
[67] S.T. Hu, *Theory of Retracts*, Wayne State University Press (1965).
[68] W. Hurewicz and H. Wallman, *Dimension Theory*, Princeton University Press (1941).
[69] D.S. Kahn, *An example in Čech cohomology spaces*, Proc. Amer. Math. Soc. **16** (1965), 584.
[70] Y. Kodama, *Note on an absolute neighborhood extensor for metric spaces*, J. Math. Soc. Japan **8** (1956), 206–215.
[71] Y. Kodama, *On a problem of Alexandroff concerning the dimension of product spaces I spaces*, J. Math. Soc. Japan **10** (1958), 380–404.
[72] Y. Kodama, *Test spaces for homological dimensions spaces*, Duke Math. J. **29** (1962), 41–50.
[73] A.N. Kolmogorov, *Über offene Abbildungen*, Ann. of Math. **38** (1937), 488–498.
[74] A. Koyama *A characterization of compacta which admit acyclic UV^{n-1}-resolutions*, Tsukuba J. Math. **20** (1996), 115–121.
[75] A. Koyama and R.B. Sher *Approximable dimension and acyclic resolutions*, Fund. Math. **152** (1997), 43–53.
[76] A. Koyama and K. Yokoi, *A unified approach of characterizations and resolutions for cohomological dimension modulo p*, Tsukuba J. Math. **18** (1994), 247–282.
[77] A. Koyama and K. Yokoi, *Cohomological dimension and acyclic resolutions*, Topology Appl. (to appear).
[78] G. Kozlowski, *Images of ANRs*, Preprint (Auburn University).
[79] G. Kozlowski and J.J. Walsh, *Cell-like mappings on 3-manifolds*, Topology **22** (1983), 147–151.
[80] V.I. Kuzminov, *Homological dimension theory*, Russian Math. Surveys **23** (1968), 1–45.

[81] H. Miller, *The Sullivan conjecture on maps from classifying spaces*, Ann. of Math. **120** (1984), 39–87.
[82] S. Mardešić and J. Segal, *Shape Theory*, North-Holland, Amsterdam (1982).
[83] W.J.R. Mitchell and D. Repovš, *The topology of cell-like mappings*, Rediconti del Seminario della Facolta di Scienze dell'Universita di Cagliari **LVIII** (1988), 265–300.
[84] T. Engel Moore, *Gromov–Hausdorff convergence to non-manifolds*, J. Geom. Anal. (to appear).
[85] J. Nagata, *Modern Dimension Theory*, Sigma Series in Pure Math., Vol. 2, Heldermann (1983).
[86] M.H.A. Newman, *A theorem on periodic transformations of spaces*, Quart. J. Math. **2** (1931), 1–8.
[87] D. Montgomery and L. Zippin, *Small subgroups of locally compact groups*, Ann. of Math. **56** (1952), 213–241.
[88] D. Montgomery and L. Zippin, *Topological Transformation Groups*, Interscience Tracts in Pure and Applied Math., Vol. 1, Interscience Publishers, New York (1955).
[89] W. Olszewski, *Completion theorem for cohomological dimensions*, Proc. Amer. Math. Soc. **123** (1995), 2261–2264.
[90] W. Olszewski, *Universal separable metrizable spaces for cohomological dimension*, Topology Appl. **61** (1995), 293–299.
[91] P. Petersen, *Gromov–Hausdorff convergence of metric spaces*, Proceedings of Symposia in Pure Math., Amer. Math. Soc. (1990).
[92] L. Pontryagin, *Sur une hypothese foundamentale de la dimension*, C. R. Acad. Sci. **190** (1930), 1105–1107 (in French).
[93] F. Raymond, *Groups of transformations on manifolds*, Proc. Amer. Math. Soc. **12** (1961), 1–7.
[94] F. Raymond and R.F. Williams, *Examples of p-adic transformation groups*, Ann. of Math. **78** (1963), 92–106.
[95] L.R. Rubin, *Characterizing cohomological dimension: The cohomological dimension of $A \cup B$*, Topology Appl. **40** (1991), 233–263.
[96] L.R. Rubin and P.J. Schapiro, *Cell-like maps onto non-compact spaces of finite cohomological dimension*, Topology Appl. **27** (1987), 221–244.
[97] L.R. Rubin and P.J. Schapiro, *Compactifications which preserve cohomological dimension*, Glasnik Matematički **28** (1993), 155–165.
[98] I.A. Shvedov, *On some general properties of dimension*, Soviet Math. Dokl. (1968).
[99] K.A. Sitnikov, *Combinatorial topology of non-closed sets I*, Mat. Sbor. **34** (1954), 3–54.
[100] E.G. Skljarenko, *On the definition of cohomology dimension spaces*, Soviet Math. Dokl. **6** (1965), 478–479.
[101] P.A. Smith, *Transformations of finite period, III, Newman's Theorem*, Ann. of Math. **42** (1941), 446–458.
[102] P.A. Smith, *Periodic and nearly periodic transformations*, Lectures in Topology, R. Wilder and W. Ayres, eds, University of Michigan Press, Ann Arbor, MI (1941), 159–190.
[103] E. Spanier, *Algebraic Topology*, McGraw-Hill, New York (1966).
[104] D. Sullivan, *Geometric Topology, Part I: Localization, Periodicity, and Galois Symmetry*, MIT Press (1970).
[105] R.G. Swan, *A new method in fixed point theory*, Comment. Math. Helv. **34** (1960), 1–16.
[106] J. von Neumann, *Die Einfuhrung analytischer Parameter in topologischen Gruppen*, Ann. of Math. **34** (1933), 170–190.
[107] J.J. Walsh, *Dimension, cohomological dimension, and cell-like mappings*, Lecture Notes in Math., Vol. 870 (1981), pp. 105–118.
[108] T. Watanabe, *A note on cohomological dimension of approximable movable spaces*, Proc. Amer. Math. Soc. **123** (9) (1995), 2883–2885.
[109] J. West, *Open problems in infinite-dimensional topology*, Open Problems in Topology, North-Holland, Amsterdam (1990).
[110] G.W. Whitehead, *Elements of Homotopy Theory*, Springer, Berlin (1978).
[111] R.F. Williams, *A useful functor and three famous examples in topology*, Trans. Amer. Math. Soc. **106** (1963), 319–329.
[112] R.F. Williams, *The construction of certain 0-dimensional transformation groups*, Trans. Amer. Math. Soc. **129** (1967), 140–156.
[113] C.T. Yang, *p-Adic transformation groups*, Mich. Math. J. **7** (1960), 201–218.
[114] A.V. Zarelua, *Finitely-multiple mappings*, Dokl. Akad. Nauk SSSR **172** (1967), 775–778.

CHAPTER 10

Flows with Knotted Closed Orbits

John Franks

Department of Mathematics, Northwestern University, Evanston, IL 60208-2730, USA

Michael C. Sullivan

Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA

Contents

Abstract

We survey results concerning dynamics of flows on S^3 with special attention to the relationship between dynamical invariants and invariants of geometric topology.

HANDBOOK OF GEOMETRIC TOPOLOGY

Edited by R.J. Daverman and R.B. Sher

1. Introduction

One of the key objects of study in the field of dynamical systems is the topological structure of the solutions of ordinary differential equations. Formally we can think of an action of the real numbers **R** on a manifold and one wants to study the topology of the orbits of this action. All such orbits are either injectively immersed copies of **R** or embedded copies of S^1. The latter are called *periodic orbits* since if we consider the flow

$$\phi_t : M \to M,$$

where $t \in \mathbf{R}$ and M is the manifold on which the flow lives then a point $x \in M$ (sometimes called an initial condition) lies on a periodic orbit if and only if $\phi_{t_0}(x) = x$ for some $t_0 \in \mathbf{R}$. An orbit of a flow is compact if and only if it is periodic.

From a classical differential equations point of view (though probably not from the topologist's viewpoint) the most useful manifolds on which to investigate flows are Euclidean spaces $\mathbf{R}^n$, or perhaps the sphere S^n if it is particularly helpful to be in a compact setting.

An area where this classical setting intersects a rich topological environment is the study of flows on S^3 or $\mathbf{R}^3$ where periodic orbits can be knotted or linked. We can then ask what knot types occur as periodic orbits of a given flow (or of any flow) as well as numerous other natural questions relating the dynamics of a flow to the topology of the knots and links which occur as closed orbits of it.

2. Flows in boxes

We begin by describing a very simple but remarkably rich construction (introduced by Birman and Williams [2]) to produce examples of flows with many knotted orbits. We will see that this is much more than a method of producing examples, however. Indeed, the behavior exhibited in these examples occurs quite generally in flows on S^3.

We want to build simple examples of flows on subsets of S^3. Here is a trivial example, but one which is important as a building block. Consider the cube, or box, $B = I \times I \times I$ where $I = [-1, 1]$ which we parameterize by (x, y, s) with $x, y, s \in I$. We consider the "flow" on B whose orbits are the line segments (x, y, s), $s \in I$. We orient the orbits in the direction of decreasing s. And we specify a constant speed of 2 so an orbit will enter a box and remain for one unit of time before exiting. Formally $\Phi_t(x, y, s) = (x, y, s - 2t)$. Strictly speaking this is not a flow because the orbits will exit B after a finite amount of time (both positively and negatively).

Conceptually we think of orbits entering B at the top ($s = 1$), flowing downward and exiting at the bottom ($s = -1$). On the sides of B the orbits lie in the boundary. This is not a very interesting example as it stands, but we can construct very interesting examples by using multiple copies of B as building blocks and attaching parts of the bottom faces of one cube to the top faces of another. It is better to refer to *entering* or *exiting* faces rather than top or bottom faces, because we will want to embed collections of attached cubes in $\mathbf{R}^3$ or S^3 in complicated ways for which "top" and "bottom" may not be useful descriptions.

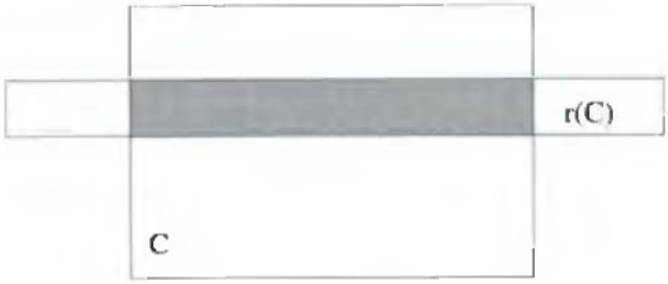


Fig. 1. The affine attaching map r.

When we attach part of the exiting face of one box to part of the entering face of another we will, of course assume that the orbit segments exiting the first at a point of attachment immediately enter the second forming a longer orbit segment crossing both boxes. We will only allow attachments of a very special form. In particular, we will make identifications using affine maps of faces which preserve the line segments in the faces parallel to the x and y axes. That is, a point $(x, y, -1)$ in the exiting region of box B_1 may be identified with its image under an affine map, $(\lambda x + a_0, \lambda^{-1} y + b_0, 1)$, in the entering region of box B_2 provided $|\lambda| > 1$ and the image under this affine map of the exiting region intersects the entering region as shown in Figure 1.

In particular the affine map, which we will denote $r(x, y)$, must satisfy $r(x, y) = (\lambda x + a_0, \lambda^{-1} y + b_0)$ where

$$\begin{aligned}
|\lambda + a_0| &> 1, \\
|-\lambda + a_0| &> 1, \\
\left|\lambda^{-1} + b_0\right| &< 1, \\
\left|-\lambda^{-1} + b_0\right| &< 1
\end{aligned}$$

so that if $C = I \times I$, the image under this map of C goes completely across C in the direction parallel to the x-axis, but in the direction parallel to the y-axis C goes completely across the image of C under this affine map.

Parts of one exiting region may be attached to more than one entering region of one or more than one box. The regions of attachment for all entering and exiting faces must be disjoint. See Figure 2. Different attachments are permitted to use different affine maps but they must all have the form above. In particular different choices of λ are permitted, but the coefficient of x (i.e., λ) must always have absolute value greater than one and, as a consequence, the coefficient of y (i.e., λ^{-1}) will have absolute value less than one.

In order to make interesting examples of flows on regions made up of boxes with these kinds of identifications we will need to make attachments that form a "cycle" of boxes, so it is possible for an orbit to leave a box B, pass through others, then return to B. Clearly this is necessary if we are to create a periodic orbit. It is not difficult to see that if we form a simple cycle of boxes like that in Figure 3 then there will be a single periodic orbit.

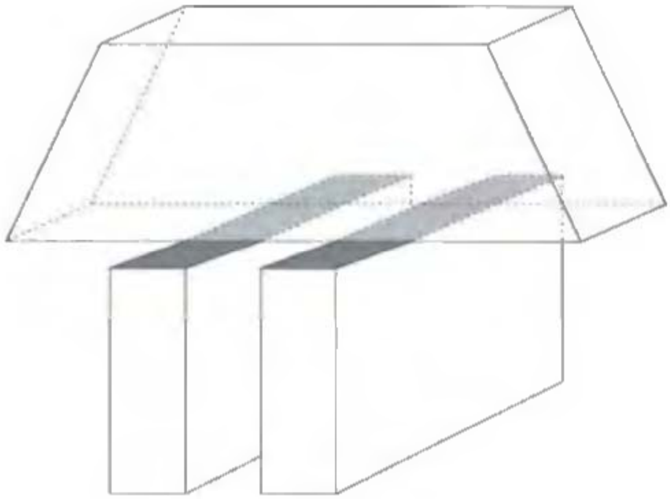


Fig. 2. Attached boxes.

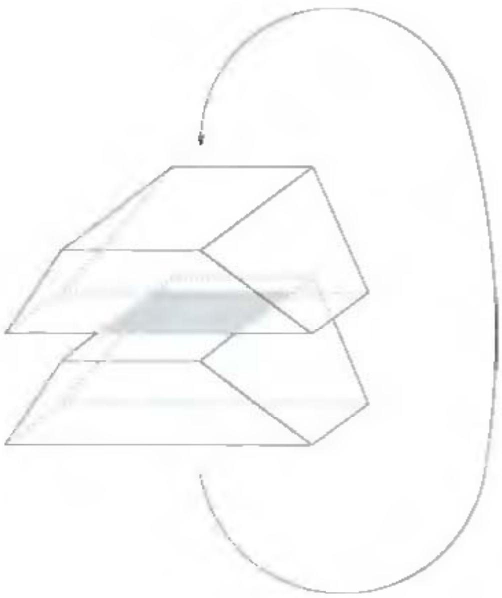


Fig. 3. Two boxes, one closed orbit.

It is also not difficult to see that every other orbit in the boxes depicted in Figure 3 will eventually exit the boxes in one direction or the other.

The easiest way to demonstrate these facts is to consider the "return map" defined on (part of) one of the entering faces. Let C be one of these entering faces and let D_0 represent the subset of points in C which will return to C when the (positively oriented) orbit through

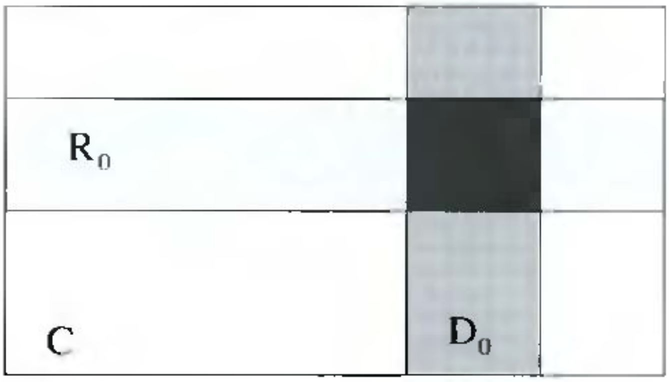


Fig. 4.

them is followed, and denote by R_0 the points to which we return. See Figure 4. Then this return map $r: D_0 \to R_0$ is an affine map of the type described above for attachments. Indeed it is just the composition of the affine maps for the attachments through which an orbit passes.

From the form of this affine map it is easy to see that r has a unique fixed point, say (x_0, y_0). Moreover, the only points (x, y) for which $r^n(x, y)$ is defined for all $n > 0$ are those with $x = x_0$ while the only points for which $r^n(x, y)$ is defined for all $n < 0$ are those with $y = y_0$. Clearly the fixed point of r corresponds to a periodic orbit and any other periodic orbit (if there were any) would correspond to a periodic point of the map r.

We want now to investigate an analogous object but one which is more complicated than a simple cycle. Assume we have n boxes $B_1, \ldots, B_n$ with multiple attachments of the type described above. We will encode the box attachments in a matrix called the transition matrix A. We define it by setting A_{ij} equal to the number of attachments of the exiting face of B_i to the entering face of B_j. The arrangement of these attachments is important, but for the moment we will ignore it. The boxes must be embedded in $\mathbf{R}^3$ in such a way as to realize the attachments and this requires that the embeddings be highly non-linear. In particular, it is perfectly legitimate for the exiting face of a box to be attached to the entering face of the same box.

Note that the transition matrix for our simple cycle in Figure 3 is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

while the one for the attached boxes shown in Figure 5 is

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Given a collection of boxes attached as described above with transition matrix A we want to investigate the collection of orbits which never exit the boxes in either direction. If

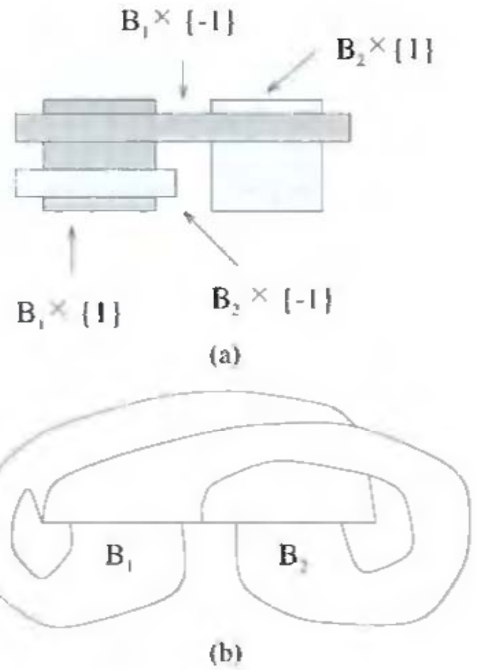


Fig. 5. Another two-box flow: (a) the first return map; (b) side view of boxes.

we denote this set of points by Ω then it is easy to see that it is a compact set in the interior of the boxes which is invariant under the flow $\psi_t : \Omega \to \Omega$ whose orbits locally (in one box) are the arcs given by holding x and y fixed and varying t in the box co-ordinates. The orbits of this flow are parameterized so that $\psi_t(x, y, t_0) = (x, y, t_0 - t)$ whenever (x, y, t_0) and $(x, y, t_0 - t)$ are in the same box.

DEFINITION 2.1. The flow $\psi_t : \Omega \to \Omega$ described above will be called the *boxed flow* associated to the collection of embedded boxes.

We will investigate this flow by considering a "return map" similar to the one described above. The name is perhaps not the best since typically points do not return under just one iteration of the map (unless the exiting face is attached to the entering face of the same box). The difference between this map and the return map discussed before is that now if there are n boxes in our construct we will consider the return map r to be defined on a subset of the union of all entering faces and having values in a different subset of the union of all entering faces. More precisely, if x is in an entering face, say of box B_i, then $r(x)$ is defined provided the (positively oriented) orbit through x exits B_i in a region of attachment and immediately enters another box, say B_j. The value of $r(x)$ is the point in the entering face of B_j which is the first point of this box on the positive orbit segment starting at x.

The subset of an entering face on which r is defined will in general have several components each of which will be a rectangle entirely crossing that face in the direction parallel to the y-axis, like D_0 in Figure 4. The image of r in each entering face will in general consist of several components each of which will be a rectangle entirely crossing that face in the direction parallel to the x-axis, like R_0 in Figure 4. Since the transition matrix A was defined by letting A_{ij} equal the number of attachments of box B_i to B_j it is clear that A_{ij} of the components of the domain of r in the entering face of B_i will be mapped by r to A_{ij} of the components of the image of r in the entering face of B_j. We will denote these domain components by $D_{ij}(k)$ and the range components by $R_{ij}(k)$, where $1 \leqslant k \leqslant A_{ij}$. Note then that the restriction $r: D_{ij}(k) \to R_{ij}(k)$ is an affine map which expands the horizontal or x component and contracts the vertical or y component.

Let D denote the union of all the domain components, i.e., the full domain of r and let

$$\Lambda = \bigcap_{n=\infty}^{\infty} r^n(D).$$

Then $r: \Lambda \to \Lambda$ is a homeomorphism (assuming Λ is non-empty).

Each point in Λ has an associated forward and backward "itinerary". This is just the sequence of rectangles $D_{ij}(k)$ through which the trajectory of the point travels. More formally the the *forward itinerary* of a point $z \in \Lambda$ is the sequence $d_0, d_1, d_2, \ldots$ where d_n is the element of the collection $\{D_{ij}(k)\}$ which contains $r^n(z)$. The *backward itinerary* of a point $z \in \Lambda$ is defined similarly as the sequence $\ldots, d_{-2}, d_{-1}, d_0$ where d_n is the element of the collection $\{D_{ij}(k)\}$ which contains $r^n(z)$. Combining the two provides the *complete itinerary* $\ldots, d_{-2}, d_{-1}, d_0, d_1, d_2, \ldots$ of the point z.

LEMMA 2.2. *Suppose $z = (x_0, y_0, 1) \in \Lambda$ is in the entering face of B_i. Then the set of $(x, y, 1) \in B_i$ with the same forward itinerary as z is*

$$\{(x_0, y, 1) \mid y \in [-1, 1]\},$$

i.e., it is the interval in the entering face of B_i which passes through z and is parallel to the y-axis. Likewise the set of $(x, y, 1) \in B_i$ with the same backward itinerary as z is $\{(x, y_0, 1) \mid x \in [-1, 1]\}$. Moreover, if two points in Λ have the same forward (backward) itinerary then the distance between the images of these points under Ψ_t tends to 0 as $t \to \infty$ ($t \to -\infty$). These line segments with the same forward (resp. backward) itinerary are called local stable manifolds, (*resp.* local unstable manifolds).

PROOF. The affine attaching maps described above contract each line segment parallel to the y-axis by a factor of λ^{-1} and expand each line segment parallel to the x-axis by a factor of λ. It follows that if $z_1 = (x_0, y, 1)$ the distance between $r^n(z)$ and $r^n(z_1)$ is $\lambda^{-n}|y - y_0|$. Hence $r^n(z_1)$ is defined for all $n \geqslant 0$ and z and z_1 have the same forward itinerary.

Conversely if $z_2 = (x, y, 1)$ is in the entering face of B_i and has the same forward itinerary as z then the distance between $r^n(z)$ and $r^n(z_2)$ is greater than $\lambda^n|x - x_0|$. Since this distance must be bounded independent of n we can conclude that $x = x_0$.

From the fact that $\|r^n(z) - r^n(z_1)\| \leqslant \lambda^{-n}|y - y_0|$ if follows that $\|\Psi(z) - \Psi(z_1)\|$ tends to 0 as $t \to \infty$.

A similar proof shows that points with backward itinerary equal to that of z are the interval specified and that the distance between such points tends to zero under Ψ_t as $t \to -\infty$. □

We can also start with a possible itinerary and show that there are points which realize it. An *allowable sequence* $\dots, d_{-2}, d_{-1}, d_0, d_1, d_2, \dots$ of elements of the collection $\{D_{ij}(k)\}$ is one for which $r(d_n)$ intersects d_{n+1} for all $n \in \mathbf{Z}$. Forward and backward allowable sequences are defined similarly.

Clearly a necessary condition for a sequence to be a complete itinerary is that it be an allowable sequence, since if $r(d_n) \cap d_{n+1} = \emptyset$ there are no points whose trajectories enter one box through d_n and the next box through d_{n+1}. The next lemma asserts that this is also a sufficient condition.

LEMMA 2.3. *Suppose that* $\dots, d_{-2}, d_{-1}, d_0, d_1, d_2, \dots$ *is an allowable sequence of elements of the collection* $\{D_{ij}(k)\}$ *with* $d_0 \subset B_i$. *Then the set of points in the entering face of* B_i *with forward itinerary* $d_0, d_1, d_2, \dots$ *is non-empty and consists of the line segment* $\{(x_0, y, 1) \mid y \in [-1, 1]\}$, *for some fixed* x_0. *Similarly the set of points in the entering face of* B_i *with backward itinerary* $\dots, d_{-2}, d_{-1}, d_0$ *is a line segment of the form* $\{(x, y_0, 1) \mid x \in [-1, 1]\}$ *for some* y_0. *And the set of points in the entering face of* B_i *with complete itinerary* $\dots, d_{-2}, d_{-1}, d_0, d_1, d_2, \dots$ *is the single point* $(x_0, y_0, 1)$.

PROOF. By induction on n it is not difficult to see that

$$W_n = \bigcap_{i=0}^{n} r^{-i}(d_i) = \{z \mid r^i(z) \in d_i,\ 0 \leqslant i \leqslant n\}$$

is a rectangle in d_0 of the form $[a, b] \times [-1, 1]$ whose width $b - a$ is $2\lambda^{-n-1}$. (The width of d_0 is $2\lambda^{-1}$.) It then follows that the set of points in the entering face of B_i with forward itinerary $d_0, d_1, d_2, \dots$ is $\bigcap_{n=0}^{\infty} W_n$ which is a line segment of the form $\{(x_0, y, 1) \mid y \in [-1, 1]\}$. A similar argument shows that the set of points with backward itinerary $\dots, d_{-2}, d_{-1}, d_0$ is a line segment of the form $\{(x, y_0, 1) \mid x \in [-1, 1]\}$ for some y_0. The set of points in the entering face of B_i with complete itinerary $\dots, d_{-2}, d_{-1}, d_0, d_1, d_2, \dots$ must be the intersection of these two line segments, namely the point $(x_0, y_0, 1)$. □

Clearly the itinerary sequences associated to a point in Λ give a great deal of information about the point. It has proven extremely valuable to abstract this concept and consider the "symbolic dynamics" associated with a transition matrix A which we describe in the next section.

It is interesting that (as we shall see) a complete topological description of Λ and the homeomorphism r depends only on the matrix A! As a consequence the topological type of the set of orbits which remain always in the boxes depends only on this matrix. But it is important to note that while A determines the abstract topological type of this space it says very little about the embedding of this space in $\mathbf{R}^3$ or S^3. It is really this embedding that is the subject of this article. And just as S^1 is a rather simple topological object with a very rich class of embeddings in S^3, it is also the case that as abstract topological objects the flows we have described above are well understood, but their embeddings in S^3 are far from understood.

The flows described in this section may seem to be of a rather special nature, and to some extent that is true. Nevertheless, the exhibit very typical behavior. In [12], for example, it is shown that for any smooth flow on $\mathbf{R}^3$ with a compact invariant set having positive topological entropy there is an invariant subset on which the flow is qualitatively equivalent to a boxed flow in a sense we define in the next section (topological equivalence). A definition of topological entropy can be found in [20]. It is a numerical invariant of the topological complexity of a flow.

3. Abstract symbolic dynamics

In this section we give an abstract "symbolic" description of a class of flows and then show that any boxed flow is topologically equivalent to such a flow. Of course, we must first specify our notion of equivalence.

DEFINITION 3.1. Two flows $\phi_t : X \to X$ and $\psi_t : Y \to Y$ are called *topologically equivalent* provided there is a homeomorphism $H : X \to Y$ carrying orbits of ϕ_t onto orbits of ψ_t. The flows are called *topologically conjugate* provided for every t and every $x \in X$ the equation $H(\phi_t(x)) = \psi_t(H(x))$ is satisfied.

Both of these are equivalence relations. From a topological point of view topological equivalence is the more appropriate notion and we will focus largely on it. Topological conjugacy is generally too rigid for our purposes. For example, the numerical value of the period of each periodic orbit is easily seen to be an invariant of topological conjugacy, from which it follows that there are uncountably many different equivalence classes.

Let A be an $n \times n$ matrix of non-negative integers. We construct a finite graph Γ with oriented edges as follows. Γ has n vertices numbered 1 to n and Γ has A_{ij} oriented edges running from vertex i to vertex j. Each edge is assumed to have unit length. We will denote these oriented edges by $e_{ij}(k)$, $1 \leqslant k \leqslant A_{ij}$, and denote by E the set of all edges.

We want to consider the set of bi-infinite sequences of edges and give it a topology. If the finite set E is given the discrete topology and we give the product topology to

$$\prod_{i=-\infty}^{\infty} E = E^{\mathbf{Z}}$$

then this space is compact, totally disconnected and perfect (assuming E has more than one element) and hence can be shown to be homeomorphic to a Cantor set.

We are interested in a subset of this space of bi-infinite sequences of edges. We will denote by Σ_A the subset of elements $\{a_i\}_{i \in \mathbf{Z}}$ with the property that for each i the edge a_i of Γ ends in the vertex where a_{i+1} begins. In other words $a_i = e_{pq}(k)$ and $a_{i+1} = e_{qr}(k')$ for some choices of p, q, r, k and k'. Thus, subset Σ_A consists of precisely those sequences of oriented edges which could be traced out by an infinite continuous path on the graph which always respected the orientation.

The fact that E is a set of edges of a graph is only of heuristic importance and we could equally well consider Σ_A as bi-infinite sequences of "symbols" satisfying a finite collection of combinatorial rules about which symbols are allowed to follow which other symbols.

It is a straightforward exercise to show that Σ_A is a closed (and hence compact) subset of $E^{\mathbf{Z}}$.

Introducing some dynamics, we define the map

$$\sigma : \Sigma_A \to \Sigma_A$$

by $\sigma(\{a_i\}) = \{b_i\}$ where $b_i = a_{i-1}$. The map σ is called the *subshift of finite type* based on the matrix A. It is a shift because it simply shifts the bi-infinite sequences of symbols one place to the left. If any symbol were allowed to follow any other (i.e., if the matrix A is a nonzero 1×1 matrix and hence any symbol is allowed to follow any other symbol) then σ is called a *full* shift. The "finite type" part of the name refers to the fact that there are finitely many rules (encoded in the matrix A) about which symbols may follow which other symbols. It is straightforward to see that σ is continuous and we can immediately exhibit its inverse (shifting to the right), so it is a homeomorphism.

There is an alternate description of Σ_A which will be useful for us. Let Γ be the graph associated to A as described above with path metric assigning unit length to each edge. Let Σ be the space of maps from $\mathbf{R}$ to Γ which preserve arc length and take integers in $\mathbf{R}$ to vertices of Γ. We give this space the compact-open topology.

It is easy to see that the map from Σ to Σ_A obtained by associating to each path of Σ the bi-infinite sequence of edges through which it passes, is a homeomorphism. More precisely if $\alpha(t)$ is an element of Σ we define $h(\alpha(t))$ to be the sequence $\{a_i\}_{i\in\mathbf{Z}}$ with the property that for each i the edge of Γ containing $\alpha(i + 1/2)$ is a_i. Then h is a homeomorphism. Also it is clear that the shift map on Σ is given by $\sigma(\alpha)(t) = \alpha(t + 1)$, or more precisely h is a topological conjugacy from $\sigma : \Sigma_A \to \Sigma_A$ to $\sigma : \Sigma \to \Sigma$.

The most thoroughly studied subshifts of finite type are those which are *irreducible*. A subshift of finite type is called *irreducible* if the graph Γ described above has the property that given any two vertices on it there is an oriented path joining them. In terms of the matrix A this is equivalent to asserting that there are no i and j such that the ijth entry of A^n is zero for all $n > 0$.

A great deal is known about irreducible subshifts of finite type. See [18], for example. We mention only that any irreducible subshift of finite type has a dense orbit and has a dense set of periodic points.

Associated to a subshift of finite type $\sigma : \Sigma_A \to \Sigma_A$ we can construct an abstract flow. The construction we use is called the "mapping torus" by topologists, but unfortunately is called the "suspension" by dynamicists.

Let X_A be the quotient space of $\Sigma_A \times \mathbf{R}$ under the identification $(\sigma(x), s) \sim (x, s + 1)$. Equivalently, let X_A be the quotient space of $\Sigma_A \times [0, 1]$ under the identification $(x, 1) \sim (\sigma(x), 0)$. The first of these descriptions is better for describing the flow ϕ_t on X, because this flow is simply the quotient of the flow Φ_t defined on $\Sigma_A \times \mathbf{R}$ by $\Phi_t(x, s) = (x, s + t)$.

In terms of the description of the subshift of finite type as the space of maps from $\mathbf{R}$ to the graph Γ associated to the matrix A, this flow also has a nice description. Recall that the subshift map σ is defined on the space Σ of arc-length preserving paths mapping $\mathbf{R}$ to Γ which take integers to vertices by $\sigma(\alpha)(s) = \alpha(s + 1)$. If we let Λ denote the space of arc-length preserving paths mapping $\mathbf{R}$ to Γ which are consistent with the orientations of the edges of Γ and give Λ the compact-open topology, then the mapping torus flow defined above is topologically conjugate to the flow on Λ given by $\Phi_t(\alpha)(s) = \alpha(s + t)$.

Because of the analogous properties for the subshift of finite type σ, whenever the matrix A is irreducible the flow ϕ_t on X_A has a dense orbit and a dense set of periodic orbits.

PROPOSITION 3.2. *A boxed flow ψ_t with transition matrix A is topologically equivalent to the mapping torus flow of the subshift of finite type $\sigma : \Sigma_A \to \Sigma_A$.*

PROOF. In fact something stronger is true. The boxed flow ψ_t with transition matrix A as described above and the mapping torus flow ϕ_t on X_A constructed from the subshift of finite type $\sigma : \Sigma_A \to \Sigma_A$, are topologically conjugate.

Recall from Section 2 that the ij entry of the transition matrix A equals the number of components of the attaching region of the exiting face of box B_i with the entering face of box B_j. Thus if we use the matrix A to construct a graph Γ and the corresponding subshift of finite type $\sigma : \Sigma_A \to \Sigma_A$, then the vertices of Γ are in one-to-one correspondence with the boxes $\{B_i\}$ and the edges from vertex i to vertex j are in one-to-one correspondence with the components of the attaching region of the exiting face of box B_i with the entering face of box B_j.

Then if $\Psi_t : \Omega \to \Omega$ is the boxed flow, this gives us a map $h : \Omega \to \Lambda$ which is a conjugacy from the boxed flow to the mapping torus flow for the subshift of finite type. It is defined by first considering the map $j : \Omega \to \Gamma$ which assigns to the point $(x, y, s) \in B_i$ the point on the edge of Γ corresponding to the region where the orbit of (x, y, s) exits B_i and whose distance from the beginning of that edge is $(1 - s)/2$. That is, its position on that edge proportional to its distance from the entering region in B_i.

Then the conjugacy h is defined using the compact-open path space description of the mapping torus flow. To a point $z \in \Omega$ we first consider its orbit $\{\Psi_t(z)\}$ in Ω and then obtain a path in Γ. This path is arc-length preserving and hence an element of Λ the space of arc-length preserving infinite paths (which is given the compact-open topology). More precisely, $h(z) = \alpha(t)$ where $\alpha(t)$ is the path $j(\Psi_t(z))$ in Γ. It is immediate that

$$h\big(\Psi_s(z)\big) = \alpha(t + s) = \Phi_s(\alpha) = \Phi_s\big(h(z)\big)$$

so h is a conjugacy.

It is straightforward to show that h is continuous. It is only necessary to check that h is one-to-one and onto, i.e., invertible, since Ω and Λ are both compact. Given an element $\alpha(t) \in \Lambda$ we wish to find $h^{-1}(\alpha)$. We can obtain the point $\alpha(0) \in \Gamma$, and from the edge of Γ we know the B_i which must contain $h^{-1}(\alpha)$. Also, if s is the distance of $\alpha(0)$ from the start of its edge, then $h^{-1}(\alpha) = (x_0, y_0, 1 - 2s)$ for some x_0 and y_0 in the co-ordinates of the box B_i. The fact that such an x_0 and y_0 exist and are unique can be shown as follows.

The fact that the edges of G are in one-to-one correspondence with the attaching regions of the exiting and entering faces of the boxes and the fact that each such attaching region is a subset of one of the domain components $\{D_{ij}(k)\}$ means that to the path $\alpha(t)$ we can associate an allowable sequence $\ldots, d_{-2}, d_{-1}, d_0, d_1, d_2, \ldots$, where each d_i is one of the domain components $\{D_{ij}(k)\}$ and they occur in the order in which the path $\alpha(t)$ crosses the corresponding edges of Γ.

According to Lemma 2.3 there is a unique point $z_0 = (x_0, y_0, 1)$ in the entering face of B_i with this itinerary. By construction $h(z_0)(t)$ is a path in G starting at the beginning of the edge containing $\alpha(0)$ and following the same sequence of edges as α. Then $h(z_0)(s) = \alpha(0)$ so $h(z_0)(t) = \alpha(t-s)$. From this it is clear that $h(\Psi_s(z)) = \alpha(t)$ and that $\Psi_s(z)$ is the unique point with this property. □

A corollary of this result and its proof is the following.

PROPOSITION 3.3. *Orbits of a boxed flow ψ_t with transition matrix A are in one-to-one correspondence with orbits of the subshift of finite type $\sigma : \Sigma_A \to \Sigma_A$. In particular, periodic orbits of ψ_t correspond to periodic orbits of σ (or, equivalently, to periodic allowable sequences of symbols associated with Σ_A).*

Rather surprisingly, if we are only interested in classifying boxed flows up to topological equivalence there is a complete and easily computed answer given by the following theorem. We emphasize that this is only for classification up to topological equivalence; the analogous question for topological conjugacy is much more subtle (see [28] or [18]).

THEOREM 3.4 (Franks [11]). *Suppose A and B are square, non-negative, integer matrices which are irreducible and Ψ and Φ are the mapping torus flows of the corresponding subshifts of finite type. Then necessary and sufficient conditions that these flows be topologically equivalent are that*

$$\det(I - A) = \det(I - B)$$

and

$$\mathbf{Z}^n/(I - A)\mathbf{Z}^n \cong \mathbf{Z}^m/(I - B)\mathbf{Z}^m,$$

where n and m are the size of A and B respectively.

Despite the fact that as abstract topological flows we can easily classify boxed flows up to topological equivalence, many problems remain if we try to understand how they are situated in three-dimensional space. It is the pursuit of some understanding of these embeddings that the next section considers.

4. Templates

In the previous sections we described "boxed flows", a class of flows defined on subsets of $\mathbf{R}^3$ (but which could be easily extended to flows on all of $\mathbf{R}^3$). In this section we want to begin the study of how these subsets are embedded. In particular any closed orbit of a flow on $\mathbf{R}^3$ is a knot and any finite set of closed orbits is a link. One of our long range objectives is the study of the relationship between dynamics and the knot types of periodic orbits.

It turns out that the union of the boxes in a boxed flow is not as conceptually simple for describing a flow as another construct called a "template" which we now describe.

DEFINITION 4.1. The *template* associated with a boxed flow Φ_s corresponding to the union of boxes $X = \bigcup B_i$ is the quotient space L of X formed by collapsing each of the boxes B_i making up X to a rectangle by identifying any two points in B_i of the form (x, y_1, t) with (x, y_2, t). There is a well defined *semi-flow* ϕ_s induced by Φ_s because any two points which are identified are carried by Φ_s to two points which are identified for all the values of s for which $\Phi_s(x, y_i, t)$ is defined.

Note that two points in X are identified in forming L if and only if they are in the same box B_i and they have the same forward itinerary. This is because we saw in Lemma 2.2 that the line segments we are collapsing to points are precisely the local stable manifolds consisting of points with the same forward itinerary.

The semi-flow ϕ_s induced by Φ_s is only "semi" because it cannot in general be inverted, i.e., it is a semi-group action of $\mathbf{R}$ rather than an action of $\mathbf{R}$ on the compact set of points on which ϕ_s is defined for all positive s.

DEFINITION 4.2. The *dual template* associated with a boxed flow Φ_s corresponding to the union of boxes $X = \bigcup B_i$ is the template of the inverse flow. Equivalently, it is the quotient space L of X formed by collapsing each of the boxes B_i making up X to a rectangle by identifying any two points in B_i of the form (x_1, y, t) with (x_2, y, t).

It is important to note that a boxed flow, by definition, includes an embedding of its boxes in $\mathbf{R}^3$ or S^3. Thus if L is the template (or dual template) associated with a boxed flow then one can embed L in $\mathbf{R}^3$ or S^3 in such a way that each point z of L lies in the interval in $\bigcup B_i$ which is collapsed to form z. Any two embeddings of L with this property are isotopic and any of them will be called an embedding of L associated to the boxed flow.

The following result is a special case of a result of Birman and Williams [2]. It tells us that understanding the closed orbits of a boxed flow in $\mathbf{R}^3$ is equivalent to understanding the closed orbits of the associated embedded template semi-flow.

THEOREM 4.3. *There is a one-to-one correspondence between the periodic orbits of a boxed flow and the periodic orbits of the induced semi-flow on the associated embedded template. Moreover, closed orbits paired by this correspondence are isotopic as embedded circles and finite sets of closed orbits matched by this correspondence are isotopic as links (embedded finite disjoint unions of circles).*

PROOF. If $h : X \to L$ is the quotient map defining L then h is a semi-conjugacy. That is, $h \circ \Phi_t = \phi_t \circ h$ for all $t \geqslant 0$. It follows that if γ is a closed orbit of Φ_t then $h(\gamma)$ is a closed orbit of ϕ_t.

Conversely, if γ_0 is a closed orbit of ϕ_t then $h^{-1}(\gamma_0)$ is a bundle over γ_0 whose fiber is an interval. This bundle (either an annulus or a Möbius strip) is invariant under the flow Φ_t and this flow preserves and contracts the fibers. If t_0 is the period of γ then Φ_{t_0} is a contraction map of each fiber to itself and hence has a unique fixed point. The collection of these fixed points is a (unique) closed orbit γ of period t_0 for Φ which is mapped by h to γ_0.

If L has an embedding associated with the flow, then for any point z of γ_0 there is precisely one point $\alpha(z)$ in $h^{-1}(z) \cap \gamma$. Sliding the point $\alpha(z)$ along the interval $h^{-1}(z)$ to z, for all points $z \in \gamma_0$ simultaneously, defines an isotopy from γ to γ_0. □

5. Knots and links

A *knot* k is an embedding of S^1 into S^3 (or $\mathbf{R}^3$), $k : S^1 \to S^3$. We are only interested in smooth embeddings. A knot may be given an orientation or preferred direction. We will always use a flow to induce an orientation on our knots. It is a common abuse of notation to use the same symbol for a knot k and its image $k(S^1) \subset S^3$. A *link* of n components is an embedding of n disjoint copies of S^1.

Two knots k_1 and k_2 (or two links) are *equivalent* if there is an isotopy of S^3 that takes k_1 to k_2. When we talk about a knot we almost always mean its equivalence class, or *knot type*. Detecting knot equivalence is the primary goal of knot theory.

In order to publish papers about knots, knot theorists have developed the *knot diagram*. This is just a projection of a knot or link into a plane such that any crossings are transverse. The crossings are then labeled as positive or negative. See Figure 6 for the convention used here. If a knot has a diagram with no crossings then it is called an *unknot* or less formally a trivial knot.

Suppose we have a knot diagram for k and we have parameterized the planar curve. If, using polar coordinates, $\mathrm{d}\theta/\mathrm{d}t > 0$ for all t, then we say that the diagram represents a *braiding* of k or that k is in *braid form*. Figure 7 shows two knot diagrams for the *figure-8 knot*, only one of which is braided. The reader is encouraged to demonstrate the equivalence. It is well known that any smooth knot or link is equivalent to a braid [6, Chapter 10]. In fact templates themselves can be braided, meaning that all the closed orbits are braided simultaneously [12]. If a knot k has a braid presentation in which all the crossings are positive then k is called a *positive braid*. Note: there exist knots which are not positive braids, but that can be presented by diagrams with only positive crossings [27].

Though a practical algorithm for detecting knot equivalence is not known, there are many useful invariants. One of special interest to us is the *genus* of a knot or link. Every link forms the boundary of an embedded orientable surface, called a Seifert surface [6]. Abstractly we can attach a disk along each boundary component of a Seifert surface, that is along each component of the link, producing a closed surface. The genus of the link is then defined to be the minimum genus over all such closed surfaces.

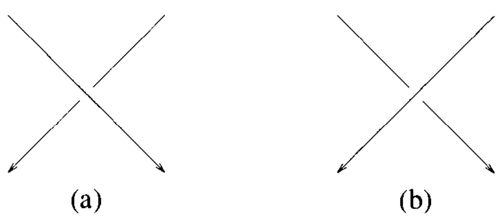


Fig. 6. (a) a positive crossing; (b) a negative crossing.

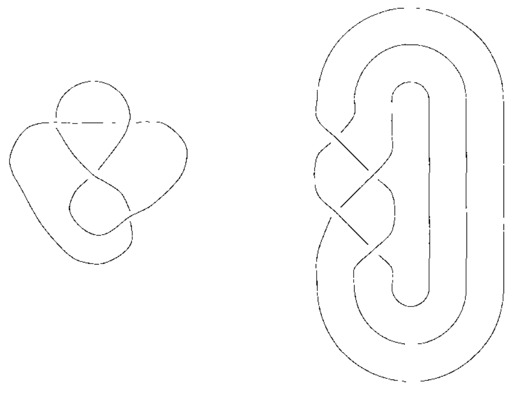


Fig. 7. The figure-8 knot.

If a knot is in braid form there are formulas which allow us to compute, or at least estimate, the genus.

PROPOSITION 5.1 (Bennequin [1]). *Let k be a knot with genus g. Suppose k has a braid presentation on s strands with* c_+ *positive crossings and* c_- *negative crossings. Then* $|c_+ - c_-| - s + 1 \leqslant 2g \leqslant |c_+ + c_-| - s + 1$.

A similar result holds for links. If the braid presentation of k is positive, then we have $2g = c - s + 1$ ($c = c_+$), a fact that was proved independently in [3].

Given a two-component link $k_1 \cup k_2$, the *linking number* of k_1 with k_2 is the sum of the signs of each crossing of k_1 under k_2 and is denoted $lk(k_1, k_2)$. The linking number is a link invariant. If the components of a link l can be separated by a 2-sphere that misses the link, then we say that l is a *split link*. For a two-component link $l = k_1 \cup k_2$, being a split link implies $lk(k_1, k_2) = 0$, though the converse is false.

The last item from knot theory we review is the notion of primeness. A knot $k \subset S^3$ is *composite* if there exists a smooth 2-sphere S^2 such that $S^2 \cap k$ is just two points p and q, and if γ is any arc on S^2 joining p to q then the knots

$$k_1 = \gamma \cup (k \cap \text{outside of } S^2) \quad \text{and}$$
$$k_2 = \gamma \cup (k \cap \text{inside of } S^2),$$

are each nontrivial (i.e., not the unknot). We call k_1 and k_2 *factors* of k and write

$$k = k_1 \# k_2.$$

We call k the *connected sum* of k_1 and k_2. If a nontrivial knot is not composite, then it is *prime*.

Figure 8 gives an example. It shows how to factor the *square knot* into two trefoils. Trefoils are prime. It was shown by Schubert [6, Chapter 5] that any knot can be factored uniquely into primes, up to order. Note: the unknot serves as a unit.

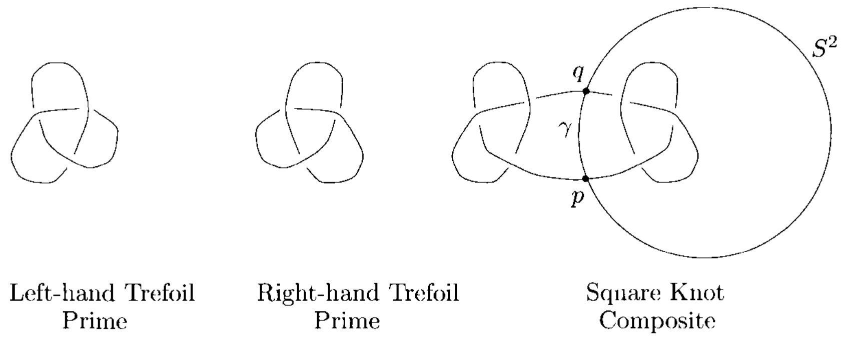


Fig. 8. The square knot is the sum of two trefoils.

6. The Lorenz template

Perhaps the simplest example of a template is the *Lorenz Template*. It is shown in Figure 9. It was developed by Williams as a naive model of the strange attractor apparently associated with the Lorenz equations, a 3×3 ODE used to study turbulent flows. See [2] and the references cited there. Although it remains unknown whether the Lorenz equations define a "Lorenz type" attractor, the existence of such attractors has been confirmed in various families of differential equations. For a brief history of this work see [20, §7.11.2] and the recent paper [8].

It is worth pointing out that boxed flows are saddle sets and not attractors. A boxed flow that would be modeled by the Lorenz template would have transition matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The reader may wish to consider how one might embed two boxes in $\mathbf{R}^3$ with ends attached as prescribed by this matrix in such a way that the associated template is essentially the one shown in Figure 9. (In this figure the corners on the edges of the template have been trimmed a bit for simplicity.)

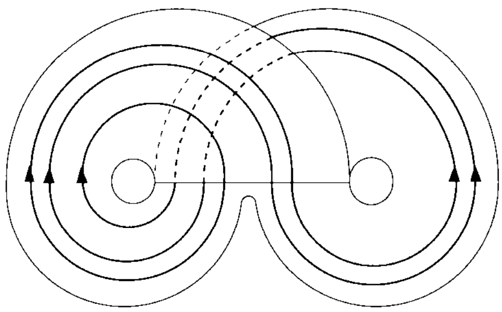


Fig. 9. The Lorenz template with a trefoil orbit shown.

Here we shall study a little of what is known about the class of knots realized as periodic orbits of the Lorenz template's semi-flow. An important symbolic tool is the use of *words* to describe orbits. We label the right band x and the left band y. Then given a staring point on the branch line, we can write down a sequence of x's and y's to describe the path of the forward orbit of our point. If the orbit is periodic we shall just record the number of symbols needed to describe one circuit. Thus the orbit shown in Figure 9 is $xxyxy$ (and turns out to be a trefoil knot). Of course, any cyclic permutation of $xxyxy$ would give the same path.

The key fact is, there is a one-to-one correspondence between the equivalence classes of finite words in two symbols under cyclic permutations and the set of periodic orbits of the Lorenz template. Similar correspondences can be set up for any template and an appropriate symbol space. This follows from the theory of subshifts of finite type and, in particular, Proposition 3.3.

Knots which occur as periodic orbits in the Lorenz template are called *Lorenz knots*.

PROPOSITION 6.1.

(1) *All Lorenz knots are positive braids with full twists,*
(2) *all torus knots (defined below) are Lorenz knots,*
(3) *all Lorenz knots are prime, and*
(4) *the only split Lorenz links have either the x or y loops as a component.*

We shall only prove (1), (2), and (4). Although Williams proved (3) directly in [29] it is now known that (1) implies (3); that is positive braids with a full twist are prime knots, as Williams himself had conjectured [7,26]. It is generally believed that there are examples of positive braids with full twists that are not Lorenz knots, but this has not to our knowledge been written up.

The proof of Proposition 6.1 (and several others) uses what is known as the *belt trick*. Consider the strip with a "loop-de-loop" as show in Figure 10. As we pull the ends apart the "loop-de-loop" turns into a full twist.

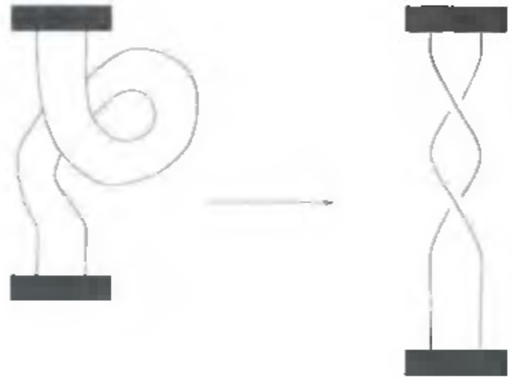


Fig. 10. The belt trick.

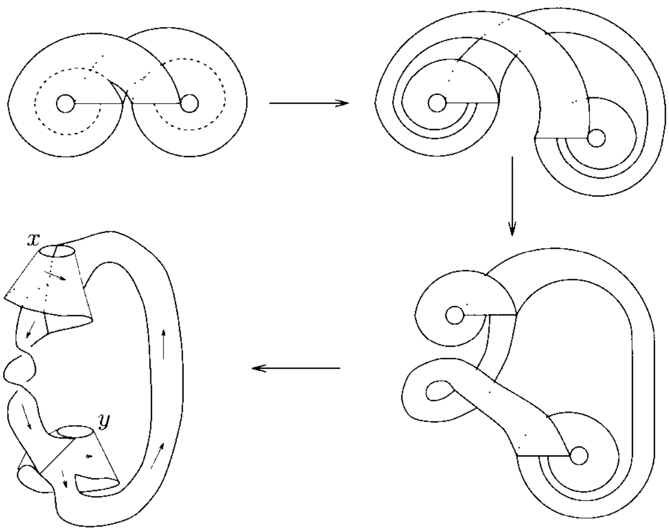


Fig. 11. Surgery on a template.

PROOF OF (1). The full twist is revealed by doing *surgery* on the template. We delete the two segments of flow lines that meet at the center of the branch line and start at the branch line. Then deform the resulting new template. But this new template has just the same periodic orbits as does the Lorenz template. The surgery and deformation are shown in Figure 11. The deformation uses the belt and in the last step a move we shall call the *lamp shade trick*. It insures that $d\theta/dt > 0$ for all orbits except the x and y orbits (which are horizontal in Figure 11. But the result holds for these two orbits trivially.

Positivity is clear as only positive crossings can be realized. □

A *torus knot* is a knot that can be represented by a closed curve on a standardly embedded torus. A torus is standardly embedded if it is the boundary of an unknotted solid torus. All torus knots can be represented by a pair of relatively prime integers (p, q). The number p gives the number of times the knot wraps around the long way (longitudinally), while q gives the number of time the knot wraps around the torus the short way (meridionally).

A torus knot can always be presented so as to have all positive or all negative crossings. We are only interested in positive torus knots. It is not hard to check that $(p, q) = (q, p)$, but otherwise the (p, q) representation is one-to-one for positive torus knots and $p, q > 0$. We shall assume $p < q$.

PROOF OF (2). In Figure 12 we show how to place a (p, q) torus knot on a torus. We do not need the whole torus; the knot can be drawn in the split strip that wraps about the torus in Figure 12. The main branch of the strip has p strands of the (p, q) knot. This part of the strip makes $n > 0$ full twists around the torus, though the figure shows only one full twist. Then the strip splits in two with a strands staying of the side of the torus facing the reader and b strands making an extra full twist around the torus. Then the two strips come back

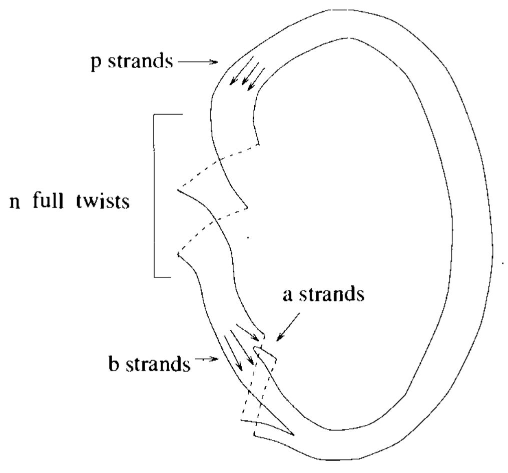


Fig. 12. Torus knots.

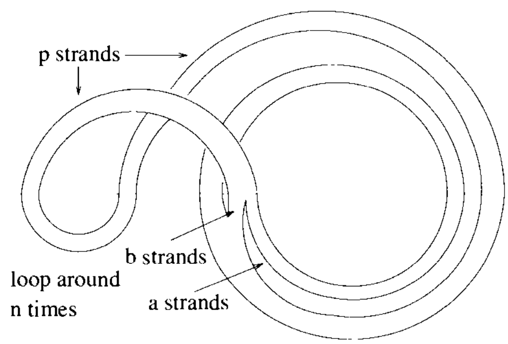


Fig. 13. Torus knots are Lorenz.

together. One sees that $q = np + b$, and that there is a single knot as long as p and q are relatively prime.

Figure 13 shows how to place the split strip onto the Lorenz template so that any torus knot can be realized as a closed orbit of the template's semi-flow. Here we have used the belt trick $n + 1$ times. Note that in this figure too, only the $n = 1$ case is shown, but the reader should be able to see how to add on extra loops of the strip about the x band of the template. □

PROOF OF (4). Any two orbits whose words involve both x's and y's must cross, and all the crossings are positive. Thus, the linking number is not zero. □

We conclude this section by stating one of the first general results that template theory has given to the study of flows and differential equations.

THEOREM 6.2 (Franks and Williams). *A smooth flow in $\mathbf{R}^3$ or S^3 that has a compact invariant set and positive topological entropy has infinitely many distinct knot types among its closed orbits.*

The *invariant* set of a flow ϕ_t on a manifold M is $\bigcap_{-\infty<t<\infty} \phi_t(M)$. We shall not define topological entropy, but only state that it a standard device for measuring "mixing". The gist of the proof is to show that any such flow must have a part of its invariant set that can be modeled by an embedded Lorenz template. Proposition 5.1 is used to show that there are closed orbits of arbitrarily high genus and thus infinity many distinct knot types can be realized as closed orbits.

7. Lorenz-like templates

By *Lorenz-like templates* we mean one of the templates depicted in Figure 14, where there are m half twists in the x branch and n half twists in the y branch. We shall denote such a template by $L(m,n)$. Thus $L(0,0)$ is the Lorenz template. Notice that $L(m,n) = L(n,m)$ via a 180° rotation. We shall at times abuse our own notation and use the symbol $L(m,n)$ to represent the set of knots and links realizable by the closed orbits in the semi-flow on $L(m,n)$.

PROPOSITION 7.1 (Williams). *Knots in $L(0,n)$ for $n \geqslant 0$ are positive braids with a full twist and hence are prime.*

PROOF. For $n = 0$ we already have this result. For $n \geqslant 2$ Figure 15 shows how to manipulate a template $L(0,n)$ to see the full twist. This is actually a simple application of the lamp shade trick used in the proof of Proposition 6.1(1). Positivity is clear. In [17] it shown that knots on $L(0, 1)$ do indeed have a full twist presentation, but there does not appear to be a presentation of the template where all knots are simultaneously presented as positive

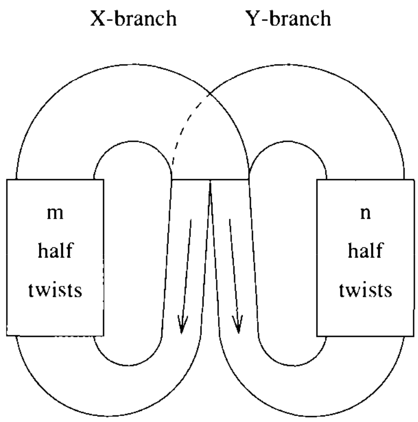


Fig. 14. Lorenz-like templates.

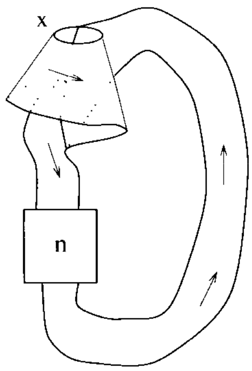


Fig. 15. At least n half twists.

braids with full twists. Also, the result in [26] shows that having a half twist in a positive braid is enough for primeness. □

PROPOSITION 7.2. *There is a lower bound on the genus of nontrivial knots in $L(0,n)$ of $g \geqslant (n-1)/2$, for $n \geqslant 0$.*

PROOF. Referring again to Figure 15, suppose we have a knot with s strands. Notice that s is also just the number of y strands. If there is only one y strand, then we have an unknot. Each half twist forces $s(s-1)/2$ crossings. Suppose $s > 1$. Thus, since all the crossings are positive, Proposition 5.1 gives $g \geqslant (n-1)/2$. □

This shows, for example, that there can be no trefoils on $L(0,n)$ for $n \geqslant 3$, since the trefoil is a genus one knot.

PROPOSITION 7.3. *As sets of knots, $L(0,n) \subset L(0,n-2)$, for all integers n.*

PROOF. The proof is pictorial. Figure 16 shows how to place $L(0,n)$ into $L(0,n-2)$. Again we make use of the belt trick, paying careful attention to the sign of the new full twist. □

PROPOSITION 7.4. *For $n < 0$, $L(0,n)$ contains composite knots.*

PROOF. Figure 17 shows a composite knot in $L(0,-1)$, found by Williams [29]. Its word is $xxyyyyxxy$ and, as the reader can check, it is a connected sum of positive and negative trefoils. Figure 18 is of a composite in $L(0,-2)$ found in [22]. The knot type is the same as before. Its word is $xxxyxxxyyxyxxyyxy$. These examples and Proposition 7.3 give the result. □

PROPOSITION 7.5. *As sets of knots $L(0,\pm 4)$ is a subset of $L(0,\pm 1)$ (respectively).*

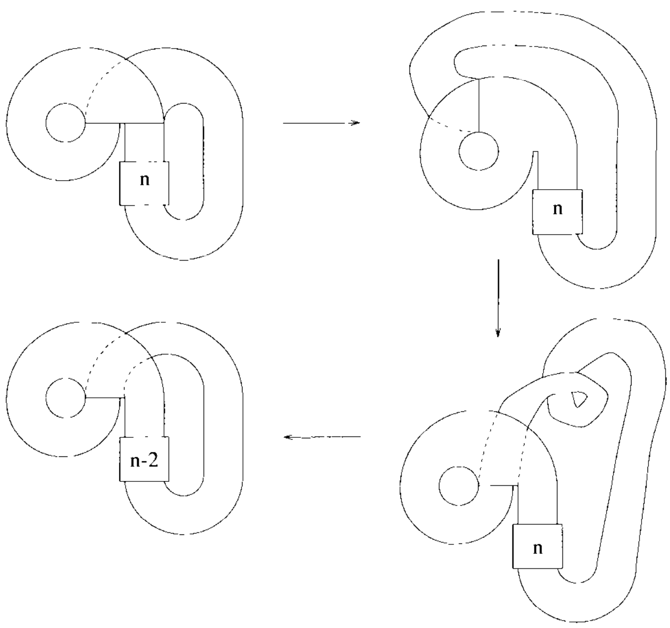


Fig. 16. $L(0, n)$ is in $L(0, n-2)$.

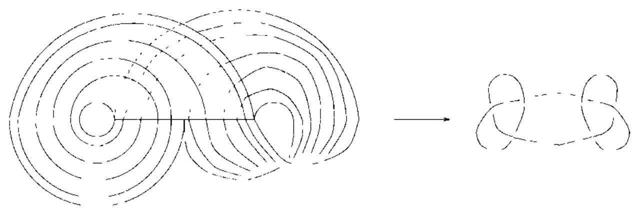


Fig. 17. A composite knot in $L(0, -1)$.

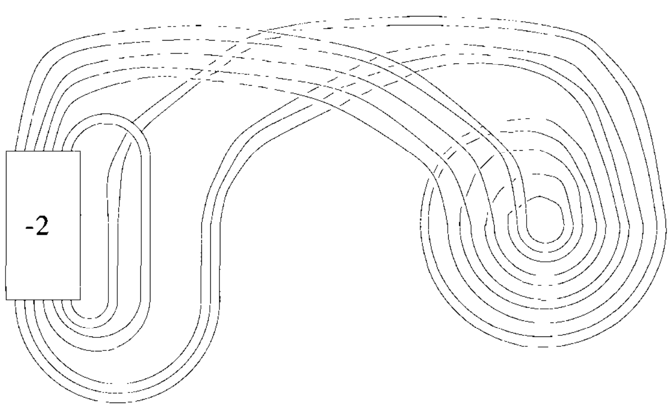


Fig. 18. A composite knot in $L(-2, 0)$.

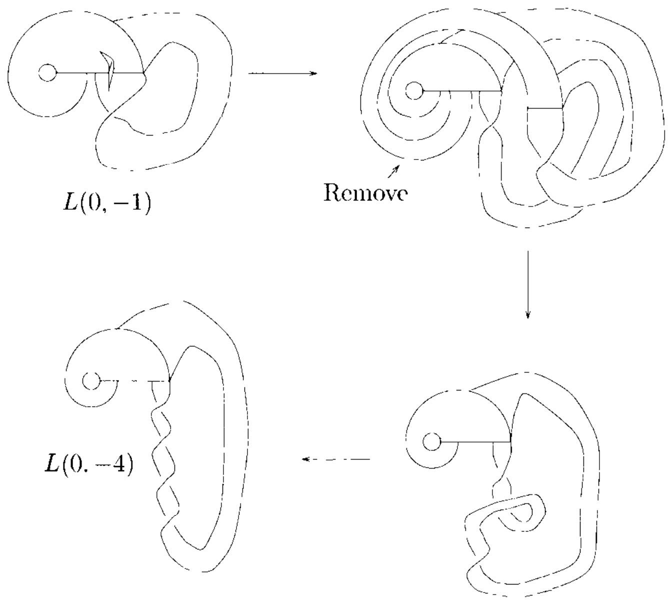


Fig. 19. $L(0, -4)$ as a subtemplate of $L(0, -1)$.

PROOF. Figure 19 gives the proof for the minus case. The surgery shown starts by cutting along the y orbit. This changes the invariant part of the semi-flow, but the knot type of the new y orbit is still that of the unknot. The linking number of the new y orbit with another closed orbit will be twice the linking number of the original y with the other orbit. So, the links have changed, but only a little. Any link in $L(0, -4)$ that does not use the y orbit is isotopic to a link in $L(0, -1)$. The plus case is similar.

The proof goes back to when one of us, as a seventh grader, was shown that when you cut a Möbius band down the middle you get a strip with four half twists in it. □

The following dichotomy has emerged. In the case $n \geqslant 0$ the templates are *positive*, that is all closed orbits are positive braids, while if $n < 0$ the templates are *mixed*, that is they have knots which are not positive braids. In the former case all the knots are prime, while in the latter case there are composite knots. In the next section we will see that this dichotomy is even stronger.

8. Universal templates

A *universal template* is one which contains all knots and links. It was originally conjectured in [3] that such objects could not exist. However, Ghrist has recently shown that they do [13]. Here we will only outline the major steps needed to show this. More recently Ghrist has found a template that contains all other templates as subtemplates.

Step 1. Call the template shown in Figure 20 U. Define U_n to be the n fold cover of U shown in Figure 21. Ghrist shows that for all $n > 0$, U_n is a subtemplate of U. It is worth

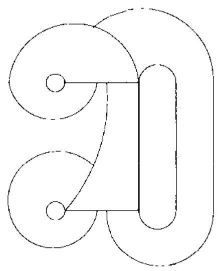


Fig. 20. The template U.

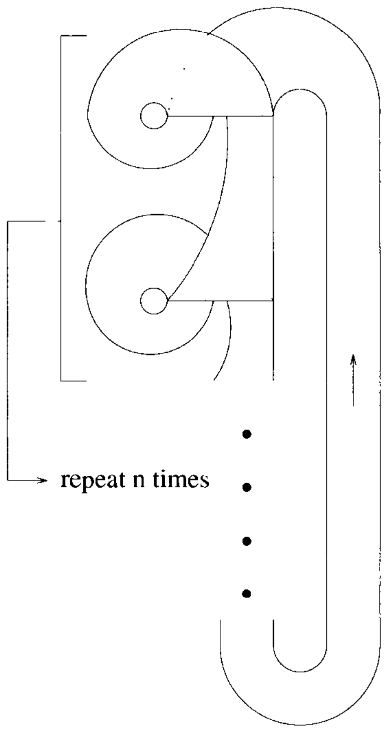


Fig. 21. The templates U_n.

noting that his method does not involve visually constructing the isotopy as above. Instead Ghrist developed a symbolic formalism for orbit-to-orbit template embeddings to show that each U_n lives in U.

Step 2. For every braid b there in an n such that $b \in U_n$. In Figure 22 we show how a negative crossing between the second and third strand can be realized on U_n. The positive crossing case is similar. By induction one can show that any braid with a single crossing is realized by a closed orbit on some U_n. We can thus construct b on a U_n by concatenation.

Fig. 22. Any crossing can be realized on U_n.

In [23] it is shown that U is a subtemplate of $L(0, -2)$. This, together with Propositions 7.3 and 7.5, shows that all the templates $L(0, n)$ for $n < 0$ are universal. While there are examples of positive templates with composite knots, e.g., $L(n, m)$ for $n, m > 1$, there always appears to be a limit on the number of prime factors. It is conjectured to be two for the example just cited. For examples (besides the Lorenz template) where such bounds have been confirmed, see [24].

CONJECTURE 8.1. *If T is a positive template, there is a number n such all periodic orbits of T have n or fewer prime factors.*

Holmes and Ghrist [14] have found differential equations and an open set of parameter values in which U arises. There are even electrical circuits governed by these equations. On the other hand, Holmes [16] has pointed out that positive templates arise naturally in the dynamics of forced oscillators, provided the hypothesis of hyperbolicity holds.

We remark that Ghrist's isotopic embedding of U_n in U is quite convoluted. For the figure-8 knot the number of strands in Ghrist's presentation runs into the millions. This is of some comfort to Sullivan and Williams, who had looked high and low for the figure-8 knot in this template without success. Still, it would be interesting to find minimum representations of a given knot or link in U. Anyone care to try?

The study of knotted periodic orbits and related phenomena is a growing and exciting field. The reader wanting to learn more may wish to consult the recently published book, *Knots and Links in Three-Dimensional Flows* [15].

References

[1] D. Bennequin, *Entrelacement et equations de Pfaff*, Asterisque **107–108** (1983), 87–161.

[2] J. Birman and R. Williams, *Knotted periodic orbits in dynamical systems I: Lorenz knots*, Topology **22** (1983), 47–82.

[3] J. Birman and R. Williams, *Knotted periodic orbits in dynamical systems II: Knot holders for fibered knots*, Contemp. Math. **20** (1983), 1–59.

[4] R. Bowen, *One-dimensional hyperbolic sets for flows*, J. Differential Equations **12** (1972), 173–179.

[5] R. Bowen and J. Franks, *Homology for zero dimensional basic sets*, Ann. of Math. **106** (1977), 73–92.

[6] G. Burde and H. Zieschang, *Knots*, Walter de Gruyter, New York (1985).

[7] P. Cromwell, *Positive braids are visually prime*, Proc. London Math. Soc. **67** (3) (1993), 384–424.

[8] F. Dumortier, H. Kokubu and H. Oka, *A degenerate singularity generating geometric Lorenz attractors*, Ergodic Theory Dynamical Systems **15** (5) (1995), 833–856.

[9] J. Franks, *Homology and Dynamical Systems*, Amer. Math. Soc. (1980).
[10] J. Franks, *Symbolic dynamics in flows on three-manifolds*, Trans. Amer. Math. Soc. **279** (1) (1983), 231–236.
[11] J. Franks, *Flow equivalence of subshifts of finite type*, Ergodic Theory Dynamical Systems **4** (1984), 53–66.
[12] J. Franks and R. Williams, *Entropy and knots*, Trans. Amer. Math. Soc. **291** (1985), 241–253.
[13] R.W. Ghrist, *Branched two-manifolds supporting all links*, Topology **36** (2) (1997), 423–438.
[14] R.W. Ghrist and P. Holmes, *An ODE whose solutions contain all knots and links*, to appear in *Int. J. Bifurcation and Chaos* (1996).
[15] R.W. Ghrist, P. Holmes and M. Sullivan, *Knots and Links in Three-Dimensional Flows*, Lecture Notes in Math., Vol. 1654, Springer, Berlin (1997).
[16] P. Holmes, *Knot and orbit genealogies in nonlinear oscillators*, New Directions in Dynamical Systems, T. Bedford and J. Swift, eds, London Math. Soc. Lecture Notes, Vol. 127, Cambridge University Press (1988), 150–191.
[17] P. Holmes and R. Williams, *Knotted periodic orbits in suspensions of Smale's horseshoe: torus knots and bifurcation sequences*, Arch. Rational Mech. Anal. **90** (1985), 115–194.
[18] D. Lind and B. Marcus, *Symbolic Dynamics and Coding*, Cambridge Univ. Press (1995).
[19] B. Parry and D. Sullivan, *A topological invariant of flows on 1-dimensional spaces*, Topology **14** (1975), 297–299.
[20] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press Inc., Boca Raton, FL (1994).
[21] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. **73** (1967), 797–817.
[22] M. Sullivan, *Prime decomposition of knots in Lorenz-like templates*, J. Knot Theory Ramifications **2** (4) (1993), 453–462.
[23] M. Sullivan, *Composite knots in the figure-8 knot complement can have any number of prime factors*, Topology Appl. **55** (1994), 261–272.
[24] M. Sullivan, *The prime decomposition of knotted periodic orbits in dynamical systems*, J. Knot Theory Ramifications **3** (1) (1994), 83–120.
[25] M. Sullivan, *A zeta-function for flows with positive templates*, Topology Appl. **66** (1995), 199–213.
[26] M. Sullivan, *Positive braids with a half twist are prime*, J. Knot Theory Ramifications **6** (3) (1997), 405–415.
[27] J.M. Van Buskirk, *Positive knots have positive Conway polynomials*, Knot Theory and Manifolds, D. Rolfsen, ed., Lecture Notes in Math., Vol. 1144, Springer (1983).
[28] R. Williams, *Classification of subshifts of finite type*, Ann. of Math. **98** (1973), 120–153; *Errata* **99** (1974), 380–381.
[29] R. Williams, *Lorenz knots are prime*, Ergodic Theory Dynamical Systems **4** (1983), 147–163.

CHAPTER 11

Nielsen Fixed Point Theory*

Ross Geoghegan

State University of New York, Binghamton, NY 13902-6000, USA

Contents

*Supported in part by NSF Grant.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Introduction

Nielsen Fixed Point Theory combines the ideas of the Lefschetz Fixed Point Theorem with the fundamental group to produce a richer version of the Lefschetz Theory. Just as the Lefschetz theorem concerns the "Lefschetz number", $L(f) \in \mathbb{Z}$, associated with a suitable map $f: X \to X$, Nielsen Theory produces an invariant called the "Reidemeister trace", $R(f)$, not a number but an element of a certain free Abelian group. In fact, $R(f)$ depends on a base point v and a base path τ from v to $f(v)$, so I will need notation later which acknowledges this; but not in this Introduction.

The theory was developed during the 1930's mainly by Reidemeister [43] and his student Wecken [50]; however, some of the main ideas appeared a little earlier in the work of Nielsen [38] and that is why the theory is named in his honor.

The *fixed point set*, $\mathrm{Fix}(f)$, of a map $f: X \to X$ is $\{x \in X \mid f(x) = x\}$. Its members are the *fixed points* of f. It is well-known that when X is a finite complex, $L(f)$ is defined, and if $\mathrm{Fix}(f) = \emptyset$ then $L(f) = 0$; indeed, this is the contrapositive of the Lefschetz theorem. Since $L(f)$ is a homotopy invariant, a better statement is: *if f is homotopic to a map g such that* $\mathrm{Fix}(g) = \emptyset$ *then* $L(f) = 0$. In this form, it is reasonable to ask if the converse is true: is the vanishing of $L(f)$ sufficient to guarantee such a g? The answer is no, and analysis of what goes wrong in the "proof" leads straight to $R(f)$ and to the theorem that, under certain hypotheses, the vanishing of $R(f)$ is sufficient to give such a g. I prove this for a reasonably general case in Section 4.

Just as with $L(f)$, there are two definitions of $R(f)$ when X is a compact manifold, one geometric and the other algebraic. And just as the classical Lefschetz–Hopf theorem says that the geometric definition of $L(f)$ (in terms of fixed point indices) equals the algebraic definition (an alternating sum of traces), so a central theorem of Nielsen theory says that under suitable hypotheses the geometric definition of $R(f)$, given here in Section 1, equals the algebraic definition given in Section 2; this is proved in Section 3.

In summary, my aim in Part I (Sections 1–4) is to give a more-or-less complete account of what I have just outlined, the qualification "more-or-less" referring mainly to two intuitively reasonable Propositions 1.1 and 1.2 whose detailed proofs are too long for this article and are well written up in the textbooks [2] and [33]. This is the basic theory, and every topologist should know it.

In Part II (Sections 5–8) I give short accounts of some connections between Nielsen Fixed Point Theory and other parts of geometric topology. Such a selection is always subjective and perhaps controversial (for what is omitted), but I think the four topics covered suggest some reasons why geometric topologists should pay attention to Nielsen theory.

Not that Part II is complete! I do not do justice to the interesting new connections between Nielsen theory and low-dimensional topology. For example, I have said that the vanishing of $R(f)$ is usually sufficient for the existence of g homotopic to f with $\mathrm{Fix}(g) = \emptyset$, but the notable exception among compact n-manifolds is the case $n = 2$. The discovery of a counter example for maps on surfaces by Jiang [29] was one of the important events in the recent resurgence of this subject. Jiang used a representation in a braid group, but more robust methods involving surface topology were subsequently developed [32,52].

Another recent development barely touched on here is parametrized Nielsen Fixed Point Theory, especially the 1-parameter version involving an interval of maps or a circle of maps; see [5,6,14–18,26]. One aspect of this theory sees the n-parameter analog of $R(f)$ as lying in the nth Hochschild homology group $HH_n(\mathbb{Z}G)$ (where $G = \pi_1(X)$) and connects the subject with geometric problems obstructed by elements in the higher K-theory of $\mathbb{Z}G$. A hint of this for the "classical" case $n = 0$ can be found in Section 5.

Nielsen theory has applications to non-linear analysis but that is outside the scope of this article; see, for example [11].

The reader interested in knowing about recent research in Nielsen Fixed Point Theory should look at the conference proceedings volumes [3,10,30,47] and, especially recommended, [35] which is the most ambitious in scope. For a different view of Fixed Point Theory *circa* 1980 (not specifically Nielsen theory) see the survey article [7].

Part I. The basic theory

1. Geometric fixed point theory

The point 0 is a fixed point of every linear map $A : \mathbb{R}^n \to \mathbb{R}^n$; it is an isolated fixed point if and only if 1 is not an eigenvalue of A, if and only if $I - A$ is non-singular; in which case the sign ($= \pm 1$) of the determinant of $I - A$ is called the *fp index* of A at 0, denoted $\iota(A, 0)$. The Jordan canonical form shows that $\iota(A, 0)$ is the parity of the number of real eigenvalues > 1, counted with multiplicity, so it is a crude measure of A's "expansiveness". But to generalize $\iota(A, 0)$ to the non-linear case, one should note instead its topological meaning: $\iota(A, 0)$ is the topological degree of the map of pairs $(I - A) : (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \to (\mathbb{R}^n, \mathbb{R}^n - \{0\})$.

If M is an n-manifold and $x \in \mathring{M}$ is an isolated fixed point of the map $f : M \to M$, the *fp index* of f at x is the degree, $\iota(f, x)$, of the map of pairs

$$\left(\mathrm{id} - hfh^{-1}\right) : \left(h(V), h(V) - \{h(x)\}\right) \to \left(\mathbb{R}^n, \mathbb{R}^n - \{0\}\right),$$

where U is open in M, $h : U \to \mathbb{R}^n$ is an open embedding (a "chart"), and V is an open n-ball neighborhood of x in U such that $f(V) \subset U$.[1]

When M is compact and $\mathrm{Fix}(f) \equiv \{x_1, \ldots, x_r\} \subset \mathring{M}$ is finite, the integer $L(f) \equiv \sum_{j=1}^{r} \iota(f, x_j)$ is the *Lefschetz number* of f. It is not obvious from this definition that $L(f)$ is a homotopy invariant, nor is it clear how to define $L(f)$ when $\mathrm{Fix}(f)$ is infinite, or M is replaced by a non-manifold. I will address these matters, but for geometric motivation I had best discuss this nicest case first: M a compact manifold, and $\mathrm{Fix}(f)$ a finite subset of its interior. By grouping together the fixed points into "fp classes", and summing the fp indices of each class separately, one gets an important refinement of the integer $L(f)$ as I now explain.

---

[1] In algebraic topology, a canonical isomorphism $o : H_n(\mathbb{R}^n, \mathbb{R}^n - \{0\}) \to \mathbb{Z}$, called the *standard orientation* on $\mathbb{R}^n$, is chosen once and for all. The map (translation of $h(x)$ to 0)$\circ$inclusion provides a canonical isomorphism $H_n(h(V), h(V) - \{h(x)\}) \to H_n(\mathbb{R}^n, \mathbb{R}^n - \{0\}) \xrightarrow{o} \mathbb{Z}$. Thus both homology groups have canonical generators mapped to $1 \in \mathbb{Z}$. Interpreting $(\mathrm{id} - hfh^{-1})_*$ as an endomorphism of $\mathbb{Z}$, the *degree* of $\mathrm{id} - hfh^{-1}$ is the image of 1. It is independent of U, h, V and any particular author's choice of standard orientation.

In preparation, I give some definitions which make sense for any self-map $f : X \to X$ of any path connected, locally connected, semilocally simply connected space. Say that $x, y \in \mathrm{Fix}(f)$ are *fp equivalent* if there is a path ν in X from x to y such that the loop $\nu(f \circ \nu)^{-1}$ in X is homotopically trivial. The equivalence classes are called *fp classes*: they are both closed and open subsets of $\mathrm{Fix}(f)$ (which, in turn, is closed in X). Thus when X is compact there are only finitely many fp classes.

Pick a base point v for X and let $G = \pi_1(X, v)$. Pick a base path τ in X from v to $f(v)$. Write $\phi : G \to G$ for the homomorphism induced by f using v and τ; i.e., if ω is a loop at v, and $[\omega]$ denotes the corresponding element of G, then $\phi([\omega]) = [\tau][f \circ \omega][\tau]^{-1}$. The homomorphism $\phi : G \to G$ defines an equivalence relation on G: say $g_1, g_2 \in G$ are *semi-conjugate* (or *ϕ-conjugate*) if there exists $g \in G$ such that $g_2 = g g_1 \phi(g)^{-1}$. Write G_ϕ for the set of semiconjugacy classes; in particular, when $\phi = 1$, the identity homomorphism, G_1 is the familiar set of conjugacy classes.

There is an injective function Φ from the set of fp classes of $f : X \to X$ to the set G_ϕ; Φ maps the fp class containing x to the semiconjugacy class containing $[\mu(f \circ \mu)^{-1}\tau^{-1}]$, where μ is any path from v to x. One checks that Φ is well-defined and injective.

Returning to $f : M \to M$ with M path connected and $\mathrm{Fix}(f) = \{x_j\} \subset \mathring{M}$ finite, let the fp classes be $\mathbf{F}_1, \dots, \mathbf{F}_s$. The integer $\iota(f, \mathbf{F}_k) = \sum_{x_j \in F_k} \iota(f, x_j)$ is called the *fp index* of f at the fp class $\mathbf{F}_k$. The principal invariant of fixed point theory is obtained by combining Φ with the fp indices of the map at the fp classes: letting $\mathbb{Z}G_\phi$ denote the free Abelian group generated by the set G_ϕ of semiconjugacy classes, the *Reidemeister trace* of f with respect to the base point $v \in M$ and the path homotopy class $[\tau]$ in M is

$$R\big(f, v, [\tau]\big) \equiv \sum_{k=1}^{s} \iota(f, \mathbf{F}_k)\Phi(\mathbf{F}_k) \in \mathbb{Z}G_\phi.$$

This is a formal integer sum of semiconjugacy classes. The number of non-zero terms in this sum is called the *Nielsen number* of f, denoted by $N(f)$; in other words, $N(f)$ is the number of fixed point classes of f having non-zero fp index.

I now generalize all this to maps $f : Z \to Z$ where Z is a path connected compact Euclidean neighborhood retract (ENR) (i.e., a retract of an open subset of a Euclidean space) and $\mathrm{Fix}(f)$ is arbitrary. Since ENR's are locally simply connected, the fp classes of f have already been defined. I must define the fp index. Let $Z \subset \mathbb{R}^N$ and let $r : W \to Z$ be a retraction where W is open in $\mathbb{R}^N$. Let $\mathbf{F}$ be an fp class of the map $f : Z \to Z$, let U be an open set in Z containing $\mathbf{F}$ such that $\mathrm{cl}\, U \cap \mathrm{Fix}(f) = \mathbf{F}$, and let $V = r^{-1}(U)$. Define $d : (V, V - F) \to (\mathbb{R}^N, \mathbb{R}^N - 0)$ by $d(x) = x - fr(x)$. There is $t > 0$ such that Z lies in $B_t(0)$, the closed ball about 0 in $\mathbb{R}^N$ of radius t. Then $H_N(\mathbb{R}^N, \mathbb{R}^N - B_t(0)) \cong \mathbb{Z}$ has a chosen generator 1, as does $H_N(\mathbb{R}^N, \mathbb{R}^N - \{0\}) \cong \mathbb{Z}$. The *fp index*, $\iota(f, \mathbf{F})$ of f at $\mathbf{F}$, is the image of the generator 1 under the following composition, considered as an endomorphism of $\mathbb{Z}$:

$$H_N\big(\mathbb{R}^N, \mathbb{R}^N - B_t(0)\big) \xrightarrow{\text{inclusion}_*} H_N\big(\mathbb{R}^N, \mathbb{R}^N - \mathbf{F}\big) \xrightarrow{\text{excision}_*^{-1}} H_N(V, V - \mathbf{F})$$
$$\xrightarrow{d_*} H_N\big(\mathbb{R}^N, \mathbb{R}^N - \{0\}\big).$$

As in the special case above, the *Reidemeister trace* is

$$R\big(f, v, [\tau]\big) \equiv \sum_{\text{fp classes } \mathbf{F}} \iota(f, \mathbf{F})\Phi(\mathbf{F}) \in \mathbb{Z}G_\phi,$$

the *Lefschetz number* is $L(f) \equiv \sum_{\text{fp classes } \mathbf{F}} \iota(f, \mathbf{F}) \in \mathbb{Z}$, and the *Nielsen number* is $N(f) \equiv$ the number of fp classes $\mathbf{F}$ for which $\iota(f, \mathbf{F}) \neq 0$.

Here are two important technical properties of the Reidemeister trace.

PROPOSITION 1.1 (Homotopy Invariance). *Let $F : Z \times I \to Z$ be a homotopy, let τ be a path from v to $F_0(v)$ and let ω be the path $F(v, \cdot)$. Then $R(F_0, v, [\tau]) = R(F_1, v, [\tau\omega])$; in particular $N(F_0) = N(F_1)$.*

Let $Z \underset{s}{\overset{r}{\rightleftarrows}} Z'$ be maps. Then r restricts to a homeomorphism $\mathrm{Fix}(s \circ r) \to \mathrm{Fix}(r \circ s)$ and induces a bijection $r_\#$ from the fp classes of $s \circ r$ to the fp classes of $r \circ s$.

PROPOSITION 1.2 (Commutativity). *Let ν be a path in Z' from v' to $r(v)$ and let μ be a path in Z from v to $s(v')$. Then $r_\#(R(s \circ r, v, [\mu(s \circ \nu)])) = R(r \circ s, v', [\nu(r \circ \mu)])$.*

The proofs of these two propositions are complicated though elementary. For a complete exposition see [2, Chapters 4 and 6]. Of course, in the simple case of manifolds when $\mathrm{Fix}(s \circ r)$ is a finite subset of the interior, Proposition 1.2 is easily proved. And in case Z is a manifold and F is transverse to projection $Z \times I \to Z$, Proposition 1.1 follows from the fact that the "fixed point set of F" is a 1-manifold whose boundary is $\mathrm{Fix}(F_0) \cup \mathrm{Fix}(F_1) \subset Z \times \{0, 1\}$; see [36] or [20, p. 123] for this method.

Finally, I should add that this definition of $R(f, v, [\tau])$ can be extended to all compact ANR's – see [2, Chapter 5].

NOTE. The definition of index for arbitrary (not necessarily discrete) fixed point sets in ENR's is not found in the original German papers [43] and [50], but came later, in the work of B. O'Neill, F. Browder, J. Leray and A. Dold. See the historical notes in [2].

2. Fixed point theory in algebraic topology

Starting again, I will redefine $R(f, v, [\tau])$ and $L(f)$ as traces in the sense of linear algebra, first on finite complexes and then on finitely dominated spaces.

Let K be a finite connected CW complex and let $f : K \to K$ be a cellular map. Choose an orientation for each cell of K, thereby specifying a basis for the cellular k-chains $C_k(K)$ with $\mathbb{Z}$-coefficients. The *Lefschetz number* of f is

$$L(f) \equiv \sum_{k \geqslant 0} (-1)^k \operatorname{trace}\big(f_k : C_k(K) \to C_k(K)\big).$$

By a well-known trick [48, 4.7.6] this is equal to

$$\sum_{k\geqslant 0}(-1)^k \operatorname{trace}\big(f_*: H_k(K) \to H_k(K)\big).^2$$

It follows that $L(f)$ is a homotopy invariant; in particular it does not depend on the choice of orientations for the cells of K.

Choose a base vertex v for K and write $G \equiv \pi_1(K, v)$. Let $p:(\widetilde{K}, \tilde{v}) \to (K, v)$ denote the universal cover. For each (oriented) cell σ of K choose a cell $\tilde{\sigma}$ of $\widetilde{K}$ so that $p(\tilde{\sigma}) = \sigma$. Orient $\tilde{\sigma}$ compatibly with σ. Choose a cellular base path τ in K from v to $f(v)$. Let $\tilde{f}: \widetilde{K} \to \widetilde{K}$ denote the lift of f such that $\tilde{f}(\tilde{v}) = \tilde{\tau}(1)$ where $\tilde{\tau}$ is the lift of τ with $\tilde{\tau}(0) = \tilde{v}$; $\tilde{f}$ is a cellular map. As in Section 1, write $\phi: G \to G$ for the homomorphism induced by f using the base path τ. The usual left action of G on $\widetilde{K}$ by covering transformations makes $C_k(\widetilde{K})$ into a finitely generated free left $\mathbb{Z}G$-module which is turned into a free right $\mathbb{Z}G$-module in the usual way, namely if $c \in C_k(\widetilde{K})$ and $g \in G$, cg is the chain $(g^{-1})_\#(c)$. The oriented k-cells $\tilde{\sigma}^k$ (one for each k-cell σ^k of K) form a $\mathbb{Z}G$ basis. The chain map $\tilde{f}_k: C_k(\widetilde{K}) \to C_k(\widetilde{K})$ is not always a $\mathbb{Z}G$-module homomorphism, as it satisfies $\tilde{f}_k(\tilde{\sigma}g) = \tilde{f}_k(\tilde{\sigma})\phi(g)$. Nevertheless one can usefully represent $\tilde{f}_k$ by a square $\mathbb{Z}G$ matrix whose (j, i) entry is the $\mathbb{Z}G$-coefficient of $\tilde{\sigma}_j^k$ in the chain $\tilde{f}_k(\tilde{\sigma}_i^k)$. Write

$$R'\big(\tilde{f}, v, [\tau]\big) = \sum_{k\geqslant 0}(-1)^k \operatorname{trace}\big(\tilde{f}_k: C_k(\widetilde{K}) \to C_k(\widetilde{K})\big).$$

This element of $\mathbb{Z}G$ depends on the choice of lifts $\tilde{\sigma}^k$ and on the choice of orientations. That is not the case with its image $R(f, v, [\tau]) \equiv q(R'(\tilde{f}, v, [\tau]))$; here $q: \mathbb{Z}G \to \mathbb{Z}G_\phi$ is induced by the function $G \to G_\phi$ which sends each g to its semiconjugacy class. Explicitly:[3]

$$R\big(f, v, [\tau]\big) \equiv \sum_{k\geqslant 0}(-1)^k q\big(\operatorname{trace}\big(\tilde{f}_k: C_k(\widetilde{K}) \to C_k(\widetilde{K})\big)\big) \in \mathbb{Z}G_\phi;$$

it is called the *Reidemeister trace*. The number of non-zero coefficients in this element of $\mathbb{Z}G_\phi$ is the *Nielsen number* of f, denoted $N(f)$. Of course, I have already given these names to apparently different quantities in Section 1. This double use of names will be cleared up in Section 3.

My immediate task is to show that $R(f, v, [\tau])$ and $N(f)$, as defined in this section, are homotopy invariants. Let $I = [0, 1]$ with the usual CW complex structure: give its vertices the orientation $+1$ and give its 1-cell the usual orientation. Give $K \times I$ the product CW complex structure and give each of its cells the product orientation. Let $F: K \times I \to K$ be

---

[2] Often, writers use rational coefficients here because $H_k(K)$ may have torsion, but what we have written has a recognized (and obvious) meaning.

[3] One should not expect an equivalent formula to be derived from $\tilde{f}_*: H_*(\widetilde{K}) \to H_*(\widetilde{K})$ since to imitate the method used for $f_*: H_*(K) \to H_*(K)$ one would need all submodules of a free $\mathbb{Z}G$-module to be free, and this is rarely true.

a cellular homotopy. Take $(v, 0)$ as base point in $K \times I$ and pick a base path τ in K from v to $F(v, 0)$. Identify $\pi_1(K \times I, (v, 0))$ with $G \equiv \pi_1(K, v)$ via the isomorphism induced by the projection $p: K \times I \to K$. Write $\phi: G \to G$ for the homomorphism

$$\pi_1\big(K \times I, (v, 0)\big) \xrightarrow{F_\#} \pi_1\big(K, F(v, 0)\big) \xrightarrow{h_{[\tau^{-1}]}} \pi_1(K, v).$$

Let $\tilde{\tau}$ be the lift of τ to $\widetilde{K}$ which starts at $\tilde{v} \in \widetilde{K}$. Let $\widetilde{F}: \widetilde{K} \times I \to \widetilde{K}$ be the lift of F which maps $(\tilde{v}, 0)$ to $\tilde{\tau}(1)$. The chain homotopy $\widetilde{D}_k: C_k(\widetilde{K}) \to C_{k+1}(\widetilde{K})$ defined on generators by $\widetilde{D}_k(\tilde{\sigma}) = (-1)^{k+1}\widetilde{F}_k(\tilde{\sigma} \times I)$ satisfies[4] $(\widetilde{D}_{k-1}\tilde{\partial}_k + \tilde{\partial}_{k+1}\widetilde{D}_k)(\tilde{\sigma}) = \widetilde{F}_{0k}(\tilde{\sigma}) - \widetilde{F}_{1k}(\tilde{\sigma})$. Clearly, $\widetilde{F}_0(\tilde{v}) = \tilde{\tau}(1)$ and $\widetilde{F}_1(\tilde{v}) = (\tilde{\tau}\tilde{\omega})(1)$, where $\omega: I \to K$ is the path $F(v, \cdot)$ in K. Since $\widetilde{D}_k(\tilde{\sigma}g) = \widetilde{D}_k(\tilde{\sigma})\phi(g)$, this leads to the matrix equation:[5] $\widetilde{D}_{k-1}\phi(\tilde{\partial}_k) + \tilde{\partial}_{k+1}\widetilde{D}_k = \widetilde{F}_{0k} - \widetilde{F}_{1k}$. Applying $\sum_{k \geqslant 0}(-1)^k q(\text{trace}(\cdot))$ to the left-hand side of this equation gives zero. Thus I have proved the first part of the following proposition; I leave the parts about the Nielsen number to the reader:

PROPOSITION 2.1 (Homotopy invariance). $R(F_0, v, [\tau]) = R(F_1, v, [\tau\omega])$; *$N(f)$ does not depend on v or $[\tau]$; $N(F_0) = N(F_1)$.*

For $\bar{g} \in G$ let $c_{\bar{g}}: G \to G$ denote the conjugation $c_{\bar{g}}(g) \equiv \bar{g}^{-1}g\bar{g}$. Right multiplication by $\bar{g}$ induces a bijection $G_\phi \to G_{c_{\bar{g}} \circ \phi}$ and hence an isomorphism $\rho_{\bar{g}}: \mathbb{Z}G_\phi \to \mathbb{Z}G_{c_{\bar{g}} \circ \phi}$. When $f(v) = v$ and the trivial base path at v is understood, we abbreviate $R(f, v, [\tau])$ to $R(f, v)$. Here is how an unpointed homotopy between pointed maps affects the Reidemeister trace:

COROLLARY 2.2. *Let $F_0(v) = v = F_1(v)$ and let ω be the loop $F(v, \cdot)$ in K at v. As base point preserving maps (i.e., using the trivial base path) let F_0 induce $\phi: G \to G$, so that F_1 induces $c_{[\omega]} \circ \phi$. Then $\rho_{[\omega]}(R(F_0, v)) = R(F_1, v)$.*

Now suppose the map $f: K \to K$ is not cellular. Again let v be a base vertex and τ a base path for f. The Cellular Approximation Theorem gives a homotopy $F: K \times I \to K$ with $F_0 = f$ and F_1 cellular. Define $R(f, v, [\tau]) = R(F_1, v, [\tau\omega])$, where $\omega = F(v, \cdot)$. By Proposition 2.1 and the fact that a homotopy between cellular maps is homotopic, rel the ends, to a cellular homotopy, this definition of $R(f, v, [\tau])$ is independent of F. Moreover, if $H: f \simeq \hat{f}$ is a homotopy, and $\lambda = H(v, \cdot)$ then $R(f, v, [\tau]) = R(\hat{f}, v, [\tau\lambda])$. In this sense, the Reidemeister trace depends only on the homotopy class of f.

---

[4] Let $E_{i,\varepsilon}$ be the face of the cube $I^k \equiv [-1, 1]^k$ obtained by holding the ith coordinate fixed at $\varepsilon = \pm 1$. Define the corresponding incidence numbers $[I^k : E_{i,-1}] = (-1)^i = -[I^k : E_{i,1}]$. At the level of cellular k-chains $\partial_k I^k = \sum_{i,\varepsilon}[I^k : E_{i,\varepsilon}]E_{i,\varepsilon}$. This convention leads to the $(-1)^{k+1}$ in the definition of $\widetilde{D}_k(\tilde{\sigma})$.

[5] The linear algebra is as follows. Let R be a ring, and $\phi: R \to R$ an endomorphism. If C is an R-matrix, write $\phi(C)$ for the matrix $[\phi(C_{ij})]$. Let A and B be $m \times m$ matrices over R. For $x \in R^m$, let X be the corresponding column matrix. Define $\alpha, \beta: R^m \to R^m$ by $\alpha(x) = A\phi(X)$ and $\beta(x) = BX$. Then for $r \in R$, $\alpha(xr) = \alpha(x)\phi(r)$ and $\beta(xr) = \beta(x)r$; so β is an endomorphism of the free right R-module R^m, while α is a "semiendomorphism". The map $\mathcal{B}: \beta \mapsto B$ [respectively $\mathcal{A}: \alpha \mapsto A$] defines an additive homomorphism from the endomorphisms [respectively semiendomorphisms] to the $m \times m$ matrices over R. The compositions $\alpha \circ \beta$ and $\beta \circ \alpha$ are semiendomorphisms, and $\mathcal{A}(\alpha \circ \beta) = A\phi(B)$ while $\mathcal{A}(\beta \circ \alpha) = BA$.

The path connected space X is *finitely dominated* if there exist a path connected finite CW complex K and maps $K \underset{u}{\overset{d}{\rightleftarrows}} X$ such that $d \circ u$ is homotopic to id_X. If K and X have base points v and x there is no loss of generality in assuming (as I always will) that $d(v) = x$, $u(x) = v$ and $d \circ u$ is homotopic to $\mathrm{id}_X \operatorname{rel}\{x\}$. If $f : X \to X$ is a map and τ is a base path from x to $f(x)$, define[6] $R(f, x, [\tau]) = d_\#(R(ufd, v, [u \circ \tau]))$. This extends the Reidemeister trace to self maps of finitely dominated spaces, in particular to compact ANR's. I must show it is well-defined.

For this I need the Commutativity Lemma 2.3 below; it appropriately expresses the naive idea $\mathrm{trace}(AB) = \mathrm{trace}(BA)$. Let (K, v) and (K', v') be finite path connected pointed CW complexes, let $K \underset{s}{\overset{r}{\rightleftarrows}} K'$ be maps, let ν be a base path in K' from v' to $r(v)$ and let μ be a base path in K from v to $s(v')$. Write $G = \pi_1(K, v)$ and $G' = \pi_1(K', v')$. Let $\phi = s_\# \circ r_\#$ and $\psi = r_\# \circ s_\#$. Then $r_\# : \mathbb{Z}G_\phi \to \mathbb{Z}G'_\psi$ is well-defined.[6]

LEMMA 2.3 (Commutativity). $r_\#(R(s \circ r, v, [\mu(s \circ \nu)])) = R(r \circ s, v', [\nu(r \circ \mu)])$.

PROOF. With the indicated base paths, $\tilde{s} \circ \tilde{r} = \widetilde{sr}$ and $\tilde{r} \circ \tilde{s} = \widetilde{rs}$. In the spirit of Footnote 5, the lemma boils down to noting that $r_\#(q(\mathrm{trace}\, \widetilde{R} s_\#(\widetilde{S}))) = q'(\mathrm{trace}\, r_\#(\widetilde{R}) r_\# s_\#(\widetilde{S}))$ which equals $q'(\mathrm{trace}(\widetilde{S} r_\#(\widetilde{R})))$ since $r_\# \circ s_\# = \psi$. Here, for each k, $\widetilde{R}$ is the $\mathbb{Z}G'$ matrix representing $\tilde{r}_k : C_k(\widetilde{K}) \to C_k(\widetilde{K}')$, and $\widetilde{S}$ is the $\mathbb{Z}G$ matrix representing $\tilde{s}_k : C_k(\widetilde{K}') \to C_k(\widetilde{K})$. □

PROPOSITION 2.4. *$R(f, x, [\tau])$ is well-defined.*

PROOF. Let $K' \underset{u'}{\overset{d'}{\rightleftarrows}} X$ be another finite domination. By 2.2 and 2.3

$$\begin{aligned} R\big(ufd, v, [u \circ \tau]\big) &= R\big(ud'u'fd'u'd, v, [u \circ d' \circ u' \circ \tau]\big) \\ &= (ud')_\#\big(R\big(u'dud'u'fd', v', [u' \circ \tau]\big)\big). \end{aligned}$$

Applying $d_\#$ to both sides gives the result. □

It follows that Proposition 2.1 and Corollary 2.2 hold for the extended definition; moreover if $h : X \to X'$ is a homotopy equivalence between finitely dominated spaces and if $f : X \to X$ and $f' : X' \to X'$ are maps such that $h \circ f$ is homotopic to $f' \circ h$ then $h_\#$ maps the Reidemeister trace of g to that of f'; I leave the precise statement to the reader.

When X is path connected and finitely dominated and $f : X \to X$ is a map, define $N(f)$ to be the number of non-zero coefficients in the element of $\mathbb{Z}G_\phi$ which is $R(f, x, [\tau])$.

6 If the following diagram of groups and homomorphisms commutes

$$\begin{array}{ccc} G & \xrightarrow{\ \phi\ } & G \\ {\scriptstyle \chi}\downarrow & & \downarrow{\scriptstyle \chi} \\ G' & \xrightarrow{\ \psi\ } & G' \end{array}$$

then there is an induced function $\chi_\# : G_\phi \to G'_\psi$, hence an induced homomorphism $\chi_\# : \mathbb{Z}G_\phi \to \mathbb{Z}G'_\psi$.

And define $L(f)$ to be the sum of all the coefficients in $R(f, x, [\tau])$. There is, of course, a competing definition of $L(f)$, namely $L'(f) = \sum_{k \geqslant 0}(-1)^k \operatorname{trace}(f_k : H_k(X) \to H_k(X))$, but:

PROPOSITION 2.5. $L(f) = L'(f)$.

PROOF. With $K \underset{u}{\overset{d}{\rightleftarrows}} X$ a finite domination as before, $L'(f) = L'(f \circ d \circ u) = L'(u \circ f \circ d)$. And I have already remarked that $L(f) \equiv L(u \circ f \circ d) = L'(u \circ f \circ d)$. □

NOTE. In my paper [13], where I present the material in this section, I attribute all this to Reidemeister [43] and Wecken [50], even though they wrote too early to have known about CW complexes and cellular chains. I learned it by reading their papers and adding what, by 1980, was common knowledge among topologists. Apart from the extension to finitely dominated spaces, I did not feel I was doing anything original. A somewhat different view is expressed in [22] which appeared at about the same time as [13] and includes this material. The term "Reidemeister trace" translates the German term used by Wecken [50]: in [22] it is given the name "generalized Lefschetz number". The latter term means something different in [12]. I urge a return to the original name "Reidemeister trace".

3. Equivalence of the two theories on ENR's

ENR's are topological retracts of compact polyhedra, while finitely dominated spaces are homotopy retracts of compact polyhedra. I will show that the definitions in Sections 1 and 2 agree where both make sense, on ENR's.

First, I relate fp classes to semiconjugacy classes in the general setting of a map $f : X \to X$ on a path connected, locally connected, semi-locally simply connected space. As before, write $G = \pi_1(X, v)$ and let $p : (\widetilde{X}, \tilde{v}) \to (X, v)$ be the universal cover. Let $x \in \operatorname{Fix}(f)$ and let μ be a path from v to x. Let $\tilde{x} = \tilde{\mu}(1)$, where $\tilde{\mu}(0) = \tilde{v}$. Then $\tilde{f}(\tilde{x}) = \tilde{x}g$ for some $g \in G$ (again the usual left action of G on $\tilde{X}$ is turned into a right action). Here is an exercise in covering space theory:

PROPOSITION 3.1. $g = [\mu(f \circ \mu)^{-1}\tau^{-1}] \in G$.

Thus the semiconjugacy class which Φ associates with the fp class of x can also be found by looking at how $\tilde{f}$ acts on any element of the fiber of p over x.

Next, consider a special case. Let J triangulate a compact n-manifold in $\mathbb{R}^n$, let K be a subdivision of J, and let $f' : K \to J$ be a simplicial map with finite fixed-point set, each fixed point lying in the interior of an n-simplex of K; this implies there can be at most one fixed point in each $\mathring{\sigma}^n$. Write $f : K \to K$ for the map f' considered as a cellular map on K.

PROPOSITION 3.2. *Let* $x \in \operatorname{Fix}(f) \cap \mathring{\sigma}^n$. *Define* $\varepsilon = 0$ [*resp.* 1] *when* f *is orientation preserving* [*resp. orientation reversing*] *near* x. *Then* $\iota(f, x) = (-1)^{n+\varepsilon}$.

PROOF. Without loss of generality, assume that $x = 0 \in \mathbb{R}^n$ and that on σ^n f agrees with a linear isomorphism $A : \mathbb{R}^n \to \mathbb{R}^n$ for which 1 is not an eigenvalue, and which maps σ^n onto an n-simplex τ in $\mathbb{R}^n$ (a simplex of J) where $\mathring{\sigma} \subset \mathring{\tau}$. Then A has no eigenvalue in the open interval $(0, 1)$, for otherwise A would not map $\partial\sigma$ onto $\partial\tau$. Thus there are no eigenvalues in $[0, 1]$, so $\operatorname{sign}(\det(I - A)) = \operatorname{sign}(\det(-A)) = (-1)^{n+\varepsilon}$. □

Combining Propositions 3.1 and 3.2 yields (for this special case):

PROPOSITION 3.3. *The respective definitions of $R(f, v, [\tau])$, $N(f)$ and $L(f)$ given in Sections* 1 *and* 2 *agree.*

By well-known methods of transversality, in particular by a version devised for this purpose called the Hopf Construction (see [2, p. 117]), any map $|J| \to |J|$ is homotopic to a map resembling f' with respect to a suitable K. This, together with the Commutativity and Homotopy Invariance properties stated in Sections 1 and 2, proves:

THEOREM 3.4. *For any map $f : Z \to Z$ on a compact ENR the respective definitions of $R(f, v, [\tau])$, $N(f)$ and $L(f)$ given in Sections* 1 *and* 2 *agree.*

Indeed, this theorem remains true if ENR is replaced by ANR, where the geometric definitions for ANRs are given in [2, Chapter 5].

4. Reducing the number of fixed points

Let M be a compact piecewise linear (PL) n-manifold in $\mathbb{R}^n$ and let $f : M \to M$ be a map.[7] Then f has finitely many fp classes, $N(f)$ of which have non-zero fp index. Intuition suggests it ought to be possible to find g homotopic to f such that g has exactly $N(f) \equiv N(g)$ fixed points; the idea being that whole fp classes should coalesce to isolated fixed points and that those isolated fixed points of index 0 should be removable. This intuition turns out to be correct when $n \neq 2$ but false when $n = 2$. The case $n = 1$ is elementary [2, p. 107]. The method described in this section works for all $n \geqslant 3$; see for example [2, Chapter 8] or [6]. However, to shorten the proof I will assume $n \geqslant 4$.

I need a special homotopy extension lemma. For $B \subset X$, a homotopy $H : B \times I \to X$ is *special* if $\operatorname{Fix}(H_t)$ is the same for all $t \in I$. If a map $f : X \to X$ satisfies $f(b) = H(b, 0)$ for all $b \in B$, the resulting map $(X \times \{0\}) \cup (B \times I) \to X$ is a *special partial homotopy*. The proof of the following elementary lemma is given in [27, § 2]:

LEMMA 4.1. *If X and B are ANR's with B closed in X, every special partial homotopy as above extends to a special homotopy $X \times I \to X$.*

[7] The restriction to n-manifolds in $\mathbb{R}^n$ allows us to give the main ideas with a minimum of technicalities; this case includes regular neighborhoods of polyhedra in $\mathbb{R}^n$. For an arbitrary n-manifold a few technical adjustments are needed which are made explicit in [6, § 11]. For the polyhedral case see [2, Chapter 8].

Returning to my set-up, I will assume that a preliminary homotopy has been performed so that f is PL with finitely many fixed points. Consider two of these fixed points, u and w, in the same fp class. Pick an embedded PL arc A in $\mathring{M}$ joining u and w such that $\mathring{A} \cap \mathrm{Fix}(f) = \emptyset$, and the paths $A \hookrightarrow M$ and $f|A : A \to M$ are homotopic rel ∂A. Choose a PL map $F : (M \times \{0\}) \cup (A \times I) \to M$ with $F(x, 0) = f(x)$ for all $x \in M$, $F(a, t) = f(a)$ when $a = u$ or w, and $F(a, 1) = a$ for all $a \in A$. By general position (since $n \geqslant 4$) assume $F(\mathring{A} \times (0, 1)) \cap F(\partial(A \times I)) = \emptyset$. Thus $F|(M \times \{0\}) \cup (A \times [0, 1 - \delta]) \to M$ is a special partial homotopy for any $\delta > 0$.

Let N be a regular neighborhood of A in $\mathring{M}$ with $\mathrm{Fix}(f) \cap N = \{u, w\}$. Let $\varepsilon > 0$ be such that $N_{2\varepsilon}(A) \subset \mathring{N}$ and $N_{2\varepsilon}(M)$ retracts to M; here, $N_\alpha(B)$ denotes the α-neighborhood of B in $\mathbb{R}^n$. Pick $0 < \delta < 1$ so that $d(F_{1-\delta}, \mathrm{id}_A) < \varepsilon$. Setting $\widetilde{F}(x, t) = f(x)$ when $x \in \mathrm{cl}(M - N)$ gives a special partial homotopy $\widetilde{F} : M \times \{0\} \cup ((\mathrm{cl}(M - N) \cup A) \times [0, 1 - \delta] \to M$ extending the appropriate part of F. By Lemma 4.1, this extends to a special homotopy $\widetilde{F} : M \times [0, 1 - \delta] \to M$. Let $g = \widetilde{F}_{1-\delta} : M \to M$. Pick a regular neighborhood N' of A such that $N' \subset N_\varepsilon(A)$ and $d(g|N', \mathrm{id}_{N'}) < \varepsilon$. Then $(\mathrm{id} - g)| : N' - \{u, v\} \to N_\varepsilon(0) - \{0\} \subset \mathbb{R}^n - \{0\}$. It follows that the degree of $(\mathrm{id} - g)| : \partial N' \to \mathbb{R}^n - \{0\}$ is $\iota(g, u) + \iota(g, w) = \iota(f, u) + \iota(f, w)$, because the homotopy is special and degree is a homotopy invariant.[8]

There are two cases. If $\iota(f, u) + \iota(f, w) = 0$ then $(\mathrm{id} - g)|\partial N'$ extends to a map $k : N' \to N_\varepsilon(0) - \{0\}$. Define $h : M \to \mathbb{R}^n$ by: $h = g$ outside N' and $h = \mathrm{id} - k$ on N'. Then h is PL, $\mathrm{Fix}(h) \cap N' = \emptyset$, $\mathrm{Fix}(h) = \mathrm{Fix}(f) - \{u, w\}$, $h(M) \subset M$ (because $N_{2\varepsilon}(A) \subset \mathring{M}$) and h is homotopic in M to g (because $N_{2\varepsilon}(M)$ retracts to M). In this way I "cancel" fixed points of opposite fp index in the same fp class.

The other case, $\iota(f, w) + \iota(f, w) = d \neq 0$, is handled similarly. One thinks of N' as a cone[8] with base $\partial N'$ and vertex u. Define $k : (N', N' - \{u\}) \to (N_\varepsilon(0), N_\varepsilon(0) - \{0\})$ by coning the map $(\mathrm{id} - g)|\partial N'$, and write $h = \mathrm{id} - k$ on N' and $h = g$ outside N'. Then $h : M \to M \subset \mathbb{R}^n$, h is homotopic in M to g, $\mathrm{Fix}(h) = \mathrm{Fix}(f) - \{w\}$, and $\iota(h, u) = d$.

In summary, I have proved a special case of the main theorem on reducing the number of fixed points of maps on manifolds:

THEOREM 4.2. *Let $f : M \to M$ be a map on a compact topological n-manifold, $n \neq 2$. Then f is homotopic to a map with precisely $N(f)$ fixed points.*

I have already discussed the failure of this theorem for $n = 2$ in the Introduction.

NOTE. Theorem 4.2 was proved for PL manifolds by G.H. Shi developing a less general version due to Wecken [50]. In fact Shi's result also holds for compact polyhedra of dimension $\geqslant 3$ such that all links of vertices are connected. The extension to topological manifolds was given by R.F. Brown. See [2, Chapter 8] for details and references. Theorem 4.2 was extended to compact polyhedra of dimension $\leqslant 2$ such that all links of vertices are connected, except for surfaces, in [27].

---

[8] In filling in the details one must use the fact that a regular neighborhood of a ball is a ball, of a triad of balls is a triad of balls, etc. See [6].

Part II. Some connections with other areas

5. The Euler characteristic and a problem about traces of idempotents

When X is finitely dominated the *Euler characteristic* of X, $\chi(X)$, is defined to be $\sum_{k\geqslant 0}(-1)^k \operatorname{rank} H_k(X)$. By Proposition 2.5, $\chi(X) = L(\mathrm{id}_X)$. But what is $R(\mathrm{id}_X, x)$? For guidance, consider the special case of a compact path connected ENR, Z. Every point $z \in Z$ is a fixed point of id_Z and all points are in the same fixed point class. Combined with Proposition 2.5 this gives:

PROPOSITION 5.1. *When Z is a path connected compact ENR, $R(\mathrm{id}_Z, z) = \chi(Z)[1] \in G_1$ where $[1] \in G_1$ is the trivial conjugacy class. Hence $N(\mathrm{id}_Z) = 0$ or 1 depending on whether $\chi(Z) = 0$ or $\chi(Z) \neq 0$.*

Proposition 5.1 suggests what should hold for any finitely dominated complex X:

CONJECTURE 5.2. $R(\mathrm{id}_X, x) = \chi(X)[1]$. *Equivalently,* $N(\mathrm{id}_X) = 0$ *or* 1.

This turns out to be equivalent to a well-known conjecture in algebraic K-theory called the Strong Conjecture of Bass [1, p. 156]. Let A be an idempotent $\mathbb{Z}G$-matrix and as before let $q : \mathbb{Z}G \to \mathbb{Z}G_1$ be induced by the function $G \to G_1$ sending each $g \in G$ to its conjugacy class.

STRONG BASS CONJECTURE. $q(\mathrm{trace}(A))$ *is an integer multiple of the trivial conjugacy class.*

The word "strong" is explained in the note at the end of this section.

To explain the point, I must recall the Wall finiteness obstruction and related matters. Denote the isomorphism class of a finitely generated projective $\mathbb{Z}G$-module P by $[P]$. The Abelian group $K_0(\mathbb{Z}G)$ is defined to be the free Abelian group generated by all such $[P]$ modulo the relations $[P] + [Q] = [P \oplus Q]$. Let $\{P\}$ be the corresponding element of $K_0(\mathbb{Z}G)$. There is a natural monomorphism $\mathbb{Z} \to K_0(\mathbb{Z}G)$ taking 1 to $\{\mathbb{Z}G\}$. The cokernel, $\widetilde{K}_0(\mathbb{Z}G)$, is the *reduced projective class group* of the ring $\mathbb{Z}G$. We write $\pi : K_0(\mathbb{Z}G) \to \widetilde{K}_0(\mathbb{Z}G)$ for the quotient homomorphism.

If A is an idempotent $m \times m$ $\mathbb{Z}G$-matrix, the image of $A : \mathbb{Z}G^m \to \mathbb{Z}G^m$ is a projective $\mathbb{Z}G$-module; conversely if P is a finitely generated projective $\mathbb{Z}G$-module it is a direct summand of a free module $\mathbb{Z}G^m$, and the projection of $\mathbb{Z}G^m$ onto the submodule P of $\mathbb{Z}G^m$ is represented by an idempotent matrix. So elements of $K_0(\mathbb{Z}G)$ can be regarded as equivalence classes $\{A\}$ of idempotent $\mathbb{Z}G$ matrices A. The homomorphism $A \mapsto \mathrm{trace}(A) \mapsto q(\mathrm{trace}(A))$ well-defines the[9] *K-theoretic trace* $T_0 : K_0(\mathbb{Z}G) \to \mathbb{Z}G_1$. One thinks of $T_0(\{A\})$ as a sort of "rank" of the projective module represented by A, and

[9] The Abelian group $\mathbb{Z}G_1$ is canonically isomorphic to $HH_0(\mathbb{Z}G)$, where $HH_*(\mathbb{Z}G)$ is the Hochschild homology of the ring $\mathbb{Z}G$. This K-theoretic trace is variously known as the Hattori–Stallings trace or the Dennis trace.

the Bass Conjecture says that this rank is essentially an integer and not a complicated element of $\mathbb{Z}G_1$.

If X is a finitely dominated space whose fundamental group is G there is an element $\sigma(X) \in \widetilde{K}_0(\mathbb{Z}G)$, called the Wall finiteness obstruction, which is zero if and only if X is homotopy equivalent to a finite complex; see [21, Theorem 6.8] for a discussion. Choose A so that $\pi(\{A\}) = \sigma(X)$, and consider $T_0(\{A\}) \in G_1$. If I is a $k \times k$ identity matrix then $\pi(\{A\}) = \pi(\{A \oplus I\})$ and $T_0(\{A \oplus I\}) = T_0(\{A\}) + k[1]$, so the sum of the coefficients in $T_0(\{A\})$ is not determined by $\sigma(X)$, but can be any integer. However, if $\pi(\{A\}) = \pi(\{B\})$ then $T_0(A) - T_0(B)$ can only be non-zero in the component of the trivial conjugacy class $[1] \in G_1$. The following is essentially proved in [13]:

THEOREM 5.3. *If $\pi(\{A\}) = \sigma(X)$ and A is chosen so that the sum of the coefficients in $T_0(\{A\})$ is $\chi(X)$ then $R(\mathrm{id}_X, x) = T_0(\{A\}) \equiv q(\mathrm{trace}(A))$.*

If Conjecture 5.2 holds for X then $q(\mathrm{trace}(A)) = \chi(X)[1]$, so the Strong Bass Conjecture holds. Conversely, if $q(\mathrm{trace}(A)) = n[1]$ for some integer n (as asserted by the Strong Bass Conjecture) then we have chosen A so that $n = \chi(X)$, implying that Conjecture 5.2 holds for X.

So Conjecture 5.2 is true for X if and only if the Strong Bass Conjecture holds for all idempotent matrices A such that $\pi(\{A\}) = \sigma(X)$. In particular, if $\sigma(X) = 0$, comparing the matrix A of Theorem 5.3 with the idempotent matrix 0, we have:

COROLLARY 5.4. *If X has the homotopy type of a finite complex then Conjecture 5.2 holds for X.*

Note that, by [51], Corollary 5.4 includes the case of X a compact ANR, and in particular a compact topological manifold.

The Strong Bass Conjecture is known to hold for all idempotent $\mathbb{Z}G$ matrices when G is residually finite, or when G is of type F (i.e., there exists a finite $K(G, 1)$ complex) and is either a subgroup of $\mathrm{GL}_n(\mathbb{C})$ or has rational cohomological dimension $\leqslant 2$, so Conjecture 5.2 is true for such fundamental groups; see [1,9,34,44].

I regard Conjecture 5.2 as one of the deepest unsolved problems of Nielsen Fixed Point Theory.

NOTE. The algebraic problem posed precisely by Bass was recognized, if vaguely, in [49, §4]. The Weak Bass Conjecture says that the sum of the coefficients of non-trivial conjugacy classes in $q(\mathrm{trace}(A))$ is 0; a corresponding weaker version of Conjecture 5.2 can easily be stated. The Weak Bass Conjecture was first stated in [8]. Note that Conjecture 5.2 is mis-stated in [28].

6. Gottlieb's theorem on the center of a group

An elegant application of fixed point theory to group theory was pointed out by Gottlieb [19]. It has been used many times by geometric topologists. The key lemma is:

LEMMA 6.1. *Let K be a finite connected CW complex, v a vertex of K, and $F: K \times I \to K$ a homotopy with $F_0 = F_1 = \mathrm{id}_K$. Write $G = \pi_1(K, v)$ and let the loop $F(v, \cdot)$ represent $g \in G$. If the Euler characteristic $\chi(K) \neq 0$ then $g = 1$.*

PROOF. Apply Corollary 2.2. In this case $R(F_0, v)$ and $R(F_1, v)$ both lie in G_1, and one has $\chi(X)[1] = \chi(X)[1]g = \chi(X)[g]$, hence $g = 1$. □

The elements $g \in G$ which arise, as in Lemma 6.1, from a homotopy $\mathrm{id}_K \simeq \mathrm{id}_K$ form a subgroup, $\mathcal{G}(K)$, of G called the *Gottlieb subgroup*. Lemma 6.1 says that when $\chi(K) \neq 0$, $\mathcal{G}(K) = \{1\}$. Obviously $\mathcal{G}(K)$ is a subgroup of $Z(G)$, the center of G. When K is aspherical $\mathcal{G}(K) = Z(G)$, as can be seen by elementary obstruction theory: build the homotopy $F: K \times I \to K$ for any given $g \in Z(G)$, skeleton by skeleton.

Recall that a group G has *type F* if there is a finite $K(G, 1)$ complex K. For such a group the *Euler characteristic*, $\chi(G)$, is defined to be $\chi(K)$. The above remarks and Lemma 6.1 imply:

THEOREM 6.2 (Gottlieb's theorem). *If G has type F and $\chi(G) \neq 0$ then G has trivial center.*

Subsequently, using different methods, Eckmann and Rosset, [43], enlarged the conclusion of Theorem 6.2 to: *...then G has no non-trivial Abelian normal subgroup.*

I have stated Gottlieb's theorem in the traditional way, but there is a theoretical case for stating a (possibly) more general version. A group G has *type FD* if there is a finitely dominated $K(G, 1)$ complex. As I write, it is unknown whether or not "type *FD*" implies "type F". The algebraic definition of $R(f, v)$ was extended to finitely dominated spaces in Section 2. Using the equivalence of Conjecture 5.2 and the Strong Bass Conjecture as explained in Section 5, one easily adapts the proofs of Lemma 6.1 and Theorem 6.2 to get:

THEOREM 6.3. *If G has type FD, if the Strong Bass Conjecture holds for all idempotent $\mathbb{Z}G$-matrices, and if $\chi(G) \neq 0$, then G has trivial center.*

In the context of one-parameter fixed point theory, specifically in [16], a "one-parameter Euler characteristic", $\chi_1(G)$, is defined for groups G of type F (in that context one thinks of $\chi(G)$ as "0-parameter") and the following is proved:

THEOREM 6.4. *If G has type F, if the Strong Bass Conjecture holds for all idempotent $\mathbb{Z}G$-matrices, and if $\chi_1(G; \mathbb{Q}) \neq 0$, then G has infinite cyclic center.*

The definition of χ_1 cannot be given here, but Theorem 6.4 is a "one-parameter Gottlieb theorem". The apparent need for the hypothesis about the Strong Bass Conjecture is connected with the need[10] for it in Theorem 6.3.

I end this section by pointing out a useful generalization of the Gottlieb subgroup. Let $f: (K, v) \to (K, v)$ be a cellular pointed map on a finite CW complex, inducing $\phi: G \to G$ as in Section 2. Let $F: K \times I \to K$ be a homotopy with $F_0 = F_1 = f$, and let $g \in G$ be

[10] For perfect symmetry, the hypothesis of Theorem 6.4 should be weakened from "type F" to "type *FD*". This can probably be done.

represented by the loop $F(v, \cdot)$; as F varies among possible homotopies, the elements g which arise this way form a subgroup $\mathcal{T}(f, v)$ of G called the *Jiang subgroup*. In particular, $\mathcal{T}(\mathrm{id}_K, v) = \mathcal{G}(K)$. Here is a useful theorem.

THEOREM 6.5. *Let $\mathcal{T}(f, v) = G$. Then*
(i) *$\phi(G)$ lies in the center $Z(G)$;*
(ii) *for some $\iota \in \mathbb{Z}$, $R(f, v) = \iota . \sum_{C \in G_\phi} C$;*
(iii) *$L(f) = \iota . N(f)$;*
(iv) *$L(f) = 0$ if and only if $N(f) = 0$ if and only if $R(f, v) = 0$;*
(v) *when $N(f) > 0$ (equivalently, $\iota \neq 0$) the injection Φ from the set of fp classes of f to the set G_ϕ of semiconjugacy classes is a bijection (hence G_ϕ is finite and all semiconjugacy classes occur as fp classes).*

PROOF. Clearly $\mathcal{T}(f, v)$ commutes with $\phi(G)$, and (i) follows. When $R(f, v) = 0$, (ii) is trivial; otherwise (ii) and (v) follow from Corollary 2.2 because $\rho_g R(f, v) = R(f, v)$ for every $g \in G$. Parts (iii) and (iv) follow from the definition of $L(f)$ and $N(f)$ in Sections 1 or 2. □

The hypothesis $\mathcal{T}(f, v) = G$ holds in various important cases: examples are given in [2, Chapter 7].

NOTE. I have given Gottlieb's original proof of Theorem 6.2. Better known is the proof in terms of finitely generated free $\mathbb{Z}G$-resolutions given by Stallings [49]. I like Gottlieb's proof because Lemma 6.1 holds for all finite complexes, aspherical or not. In [16], Theorem 6.4 is proved assuming only the Weak Bass Conjecture over $\mathbb{Q}$, but this requires a more sophisticated replacement for Theorem 6.3.

7. Periodic points, zeta functions and torsion

A. *Periodic point theory*

Let $f : (X, v) \to (X, v)$ be a (pointed) map, let $G = \pi_1(X, v)$ and let f induce $\phi : G \to G$. Throughout this section X is a compact ENR or a finite CW complex as appropriate (see Sections 1, 2), and when X is a finite complex v is a vertex and f is a cellular map.

A *periodic point of period m* is a fixed point of the m-fold iterate $f^m : X \to X$. Nielsen Periodic Point Theory is concerned with the question: *what are necessary and sufficient conditions for f to be homotopic to a map g having no periodic points*? Of course, just as in Section 4, one would only hope for a complete answer when X is a compact n-manifold, $n \neq 2$, or when X is a suitable compact polyhedron as in [2, Chapter 8]. Whereas a satisfactory answer was given to the analogous question for fixed points in Part I, a complete answer to the periodic point question is not known as I write.

A necessary condition for having such a g is that the sequence $\langle R(f^m)\rangle \in \prod_{m=1}^{\infty} \mathbb{Z}G_{\phi^m}$ be trivial. Hence, a necessary condition is that the sequence of Lefschetz numbers $\langle L(f^m)\rangle \in \prod_{m=1}^{\infty} \mathbb{Z}$ be trivial. For less crude but still computable sequences between these two extremes, one can consider a linear representation $\rho : G \to \mathrm{GL}_r(S)$ where S is a commutative ring (with 1) and $\rho \circ \phi = \rho$. The corresponding sequence is $\langle L(f^m, \rho)\rangle \in \prod_{m=1}^{\infty} S$, where $L(f, \rho)$, the *ρ-twisted Lefschetz number* of f, is defined to be the image of $R(f, v)$ under the homomorphism induced by ρ. In detail, ρ induces $\bar{\rho}: \mathbb{Z}G \to M_r(S)$, the ring of $r \times r$ matrices over S, and hence $\rho_* \equiv \text{trace} \circ \bar{\rho}: \mathbb{Z}G \to S$. Define $L(f, \rho) = \sum_{k \geqslant 0}(-1)^k \operatorname{trace} \rho_*(\tilde{f}_k) \in S$, where ρ_* maps $\mathbb{Z}G$-matrices to S-matrices entry by entry. Of course, for this to make sense I assume here that f is a cellular map on a finite complex. From the ENR point of view the homomorphism ρ induces a function $G_\phi \to \mathrm{GL}_r(S)$ because $\rho \circ \phi = \rho$, so $L(f, \rho)$ coalesces the components of $R(f, v) = \sum_{\text{fp classes } \mathbf{F}} \iota(f, \mathbf{F})\Phi(\mathbf{F})$ to give first an S-matrix, and then an element of S – its trace. When $S = \mathbb{Z}$, $r = 1$ and $\rho(G) = \{1\} \subset \mathbb{Z}$, then $L(f, \rho) = L(f)$.

All this can be summarized, if tautologically, as:

THEOREM 7.1. *Let $f: (X, v) \to (X, v)$ be a cellular map on a finite complex. If the sequence $\langle R(f^m, v)\rangle$ is non-trivial, or if, for any linear representation ρ of $G = \pi_1(X, v)$ as above, the sequence $\langle L(f^m, \rho)\rangle$ is non-trivial, then f is not homotopic to a periodic point free map.*

A more sophisticated view relates the sequences $\langle L(f^m, \rho)\rangle$ to torsion and to zeta functions. Let $\mathbb{Z}G^\phi[[t]]$ denote the *ϕ-twisted power series ring* over $\mathbb{Z}G$, consisting of power series $\sum_{m \geqslant 0} u_m t^m$ in a formal variable t, with $u_m \in \mathbb{Z}G$, subject to the multiplication rule $tg = \phi(g)t$ where $g \in G$. When A is a square $\mathbb{Z}G$-matrix, the matrix $I - At$ is invertible in the ring of $\mathbb{Z}G^\phi[[t]]$-matrices; its inverse is $\sum_{m \geqslant 0}(At)^m$. This simple observation turns out to underlie everything.

Recall the definition of K_1 of a ring. For any ring S (with 1), $\mathrm{GL}(S)$ denotes the direct limit of $\mathrm{GL}_1(S) \hookrightarrow \mathrm{GL}_2(S) \hookrightarrow \cdots$ where the rth inclusion takes the $r \times r$ matrix M to the $(r+1) \times (r+1)$ matrix $M \oplus [1]$. The Abelianization $\mathrm{GL}(S)/[\mathrm{GL}(S) : \mathrm{GL}(S)]$ is denoted by $K_1(S)$. The *torsion* of M is the element of $K_1(S)$ which M represents. A good reference for background is [4].

When X is a finite complex and f is cellular, let $\Delta(f, v) \in K_1(\mathbb{Z}G^\phi[[t]])$ be represented by the matrix $c(f) \equiv \prod_{k \geqslant 0}(I - \tilde{f}_k t)^{(-1)^{k+1}}$. While this matrix depends on choices (see Section 2), $\Delta(f, v)$ does not [15, Proposition 5.3]. If $\rho: G \to \mathrm{GL}_r(S)$ is as above, ρ induces a homomorphism $K_1(\mathbb{Z}G^\phi[[t]]) \to K_1(S[[t]])$. The image of $\Delta(f, v)$ is $\Delta(f, \rho)$ represented by $c(f, \rho) \equiv \prod_{k \geqslant 0}(I - \rho_*(\tilde{f}_k)t)^{(-1)^{k+1}}$. Applying the determinant map $\det: K_1(S[[t]]) \to S[[t]]$ we get

$$\det \Delta(f, \rho) = \prod_{k \geqslant 0} \det\bigl(I - \rho_*(\tilde{f}_k)t\bigr)^{(-1)^{k+1}} \in S[[t]].$$

Define the *ρ-twisted zeta function*, $\zeta(f, \rho) \in S[[t]]$, to be $\exp(\sum_{m=1}^{\infty} \frac{1}{m} L(f^m, \rho)t^m)$. Then "rationality of the zeta function" is expressed by:

THEOREM 7.2. $\det \Delta(f, \rho) = \zeta(f, \rho)$.

I will discuss the proof of this theorem later in the section. For now, the point to note is that $\langle L(f^m, \rho)\rangle$ is trivial if and only if $\zeta(f, \rho)$ is identically 1. So we have:

COROLLARY 7.3. *With hypotheses as in Theorem* 7.1, *if* $\det \Delta(f, \rho)$ *is not identically* 1 *then* f *is not homotopic to a periodic point free map.*

REMARK. $\Delta(f, v)$ is the torsion of the matrix $c(f)$, and $\Delta(f, \rho)$ is the torsion of the matrix $c(f, \rho)$. For historical reasons the torsion of $\det c(f, \rho)$ is called "Reidemeister torsion". So the zeta function $\zeta(f, \rho)$ is a Reidemeister torsion element.

All this is usually discussed in the context of the suspension (semi)flow on the mapping torus of f, as I now explain.

B. *Dynamics*

If $\mathbf{F}$ is an fp class of the map f^m, so are $f(\mathbf{F}), \ldots, f^{m-1}(\mathbf{F})$, and the corresponding *m-periodic orbit class* of f is $\bigcup_{j=0}^{m-1} f^j(\mathbf{F})$, where $f^0(\mathbf{F}) \equiv \mathbf{F}$. Note that the fp classes $f^i(\mathbf{F})$ and $f^j(\mathbf{F})$ are either identical or disjoint. One classifies m-periodic orbit classes algebraically thus: the relation $g \sim \phi(g)$ well-defines an equivalence relation on the set G_{ϕ^m}; write $G_{\phi^m}/\langle\phi\rangle$ for the set of equivalence classes. By analogy with Φ in Section 1, there is an injective function Φ' from the set of m-periodic orbit classes of f to the set $G_{\phi^m}/\langle\phi\rangle$. When X is a compact ENR, define

$$\overline{R}(f^m, v) = \sum_{\substack{m\text{-periodic} \\ \text{orbit classes } \mathbf{O}}} \iota(f^m, \mathbf{O})\Phi'(\mathbf{O}) \in \mathbb{Z}[G_{\phi^m}/\langle\phi\rangle].$$

Under the obvious quotient homomorphism $R(f^m, v)$ is taken to $\overline{R}(f^m, v)$, the effect being to coalesce the distinct fp classes of f^m which are in the same m-periodic orbit class of f, adding up the fp indices of the distinct pieces.

Let $T(f)$ be the mapping torus of f, i.e., the quotient space $X \times I/(x, 1) \sim (f(x), 0)$. Its fundamental group[11] is $\Gamma = \langle G, t \mid tgt^{-1} = \phi(g)\ \forall g \in G\rangle$; when ϕ is an isomorphism this is the semidirect product of G and $\mathbb{Z}$ over ϕ. As before, Γ_1 denotes the set of conjugacy classes in Γ, and the function $G_{\phi^m} \to \Gamma_1$ sending the semiconjugacy class of g to the conjugacy class of gt^m maps $\coprod_{m\in\mathbb{Z}} G_{\phi^m}$ surjectively onto Γ_1, and induces a bijection $\coprod_{m\in\mathbb{Z}}(G_{\phi^m}/\langle\phi\rangle) \to \Gamma_1$. Write $\Psi : \bigoplus_m \mathbb{Z}[G_{\phi^m}/\langle\phi\rangle] \to \mathbb{Z}\Gamma_1$ for the induced isomorphism, and write $\widetilde{R}(f^m, v) = \Psi(\overline{R}(f^m, v))$. Thus,

$$\widetilde{R}(f^m, v) = \sum_{\substack{m\text{-periodic} \\ \text{orbit classes } \mathbf{O}}} \iota(f^m, \mathbf{O})\Psi\Phi'(\mathbf{O}) \in \mathbb{Z}\Gamma_1;$$

$\widetilde{R}(f^m, v)$ and $\overline{R}(f^m, v)$ encode exactly the same information.

The algebraic formula for $\widetilde{R}(f^m, v)$ is worth noting. Abusing notation, let $q : \mathbb{Z}\Gamma \to \mathbb{Z}\Gamma_1$ be the quotient homomorphism induced by the function $\Gamma \to \Gamma_1$ sending each element to its conjugacy class.

[11] The use of t to be an element of Γ here and a formal variable elsewhere is deliberate and is intended to be suggestive. See §5 of [15].

PROPOSITION 7.4. $\widetilde{R}(f^m, v) = \sum_{k\geqslant 0}(-1)^k q(\text{trace}(\tilde{f}_k t)^m)$.

PROOF. $(\tilde{f^m})_k = \tilde{f}_k \cdot \phi(\tilde{f}_k) \ldots \phi^{m-1}(\tilde{f}_k)$, so in $\mathbb{Z}\Gamma$ one has $(\tilde{f^m})_k t^m = (\tilde{f}_k t)^m$. □

There is a *suspension semiflow* on $T(f)$, i.e., a continuous action of the additive semigroup of non-negative real numbers, in which each point flows "forward" round and round the mapping torus. In the most interesting case, f is a diffeomorphism of a compact manifold, and then the semiflow is smooth and extends to a *suspension flow* (a smooth action of $\mathbb{R}$). A point of period m corresponds to a *time m closed orbit* in the semiflow. One easily proves:

PROPOSITION 7.5. *Points y and z of period m give rise to freely homotopic time m closed orbits in the suspension semiflow if and only if y and z are in the same m-periodic orbit class of f.*

This is why m-periodic orbit classes are useful.

The *Lefschetz–Nielsen series* of f is the sequence $\langle \overline{R}(f^m, v)\rangle \in \prod_{m\geqslant 1} \mathbb{Z}[G_{\phi^m}/\langle\phi\rangle]$, or, equivalently, $\langle \widetilde{R}(f^m, v)\rangle \in \prod_{m\geqslant 1} \mathbb{Z}\Gamma_1$. There is a corresponding sequence of non-negative integers $\langle \overline{N}(f^m)\rangle$ where $\overline{N}(f^m)$, the *Nielsen number of time m closed orbits*, is the number of non-zero components in $\overline{R}(f^m, v)$ or $\widetilde{R}(f^m, v)$. By Proposition 7.5 we have:

PROPOSITION 7.6. *There are at least $\overline{N}(f^m)$ time m closed orbits in the suspension semiflow associated with f.*

For $\rho: G \to \mathrm{GL}_r(S)$ as above, the induced homomorphism which takes $R(f^m, v)$ to $L(f^m, \rho)$ factors through $\mathbb{Z}[G_{\phi^m}/\langle\phi\rangle]$, so the sequence $\langle L(f^m, \rho)\rangle$ is also a "linearization" of the Lefschetz–Nielsen series. In fact, there is a homomorphism

$$K_1(\mathbb{Z}G^{\phi}[[t]]) \to \prod_{m\geqslant 1} \mathbb{Z}[G_{\phi^m}/\langle\phi\rangle] \to S[[t]]$$

given explicitly in the proof of Theorem 5.13 of [15] which maps $\Delta(f, v) \mapsto \langle \overline{R}(f^m, v)\rangle \mapsto \zeta(f, \rho)$. Theorem 7.2 follows from the fact that this homomorphism equals

$$\Delta(f, v) \mapsto \Delta(f, \rho) \mapsto \det \Delta(f, \rho).$$

In fact, Theorem 7.2 is a corollary of a deeper "rationality theorem" [15, Theorems 5.6 and 5.10] which cannot be stated at the level of this article.

ADDED IN PROOF. This subject has seen important development since this chapter was written. See, for example, [23,24,39–42,45,46,48]. □

NOTE. The contents of this section are distilled from Sections 5, 6 of [15], somewhat influenced by [31]. The Lefschetz zeta function appeared in [37] and various zeta functions of $\zeta(f, \rho)$-type in [12].

8. Growth rate of Nielsen numbers

The *growth rate* of a sequence $\langle a_m \rangle$ of complex numbers is

$$\operatorname*{growth}_{m\to\infty} a_m \equiv \max\left\{1, \limsup_{m\to\infty} |a_m|^{1/m}\right\};$$

the value ∞ is permitted. When growth $a_m > 1$, $\langle a_m \rangle$ *grows exponentially*. I will describe some results about the growth rate of numbers associated with $f^m : X \to X$ as $m \to \infty$.

For any set C, the free Abelian group $\mathbb{Z}C$ is given the L^1-norm: $\|\sum_{c\in C} n(c)c\|_{\mathbb{Z}C} = \sum_c |n(c)|$; the subscript $\mathbb{Z}C$ will often be omitted. $\|R(f^m, v)\| = \sum_{\substack{\text{fixed point}\\ \text{classes } \mathbf{F}}} |\iota(f^m, \mathbf{F})|$ and $\|\overline{R}(f^m, v)\| = \sum_{\substack{m\text{-orbit}\\ \text{classes } \mathbf{O}}} |\iota(f^m, \mathbf{O})|$. Recall from previous sections the numbers $N(f^m)$ and $\overline{N}(f^m)$. Asymptotic versions of these are $N^\infty(f) \equiv \operatorname{growth}_{m\to\infty} N(f^m) = \operatorname{growth}_{m\to\infty} \overline{N}(f^m)$ and $R^\infty(f) \equiv \operatorname{growth}_{m\to\infty} \|R(f^m, v)\| = \operatorname{growth}_{m\to\infty} \|\overline{R}(f^m, v)\|$. The equalities are consequences of the inequalities $\overline{N}(f^m) \leqslant N(f^m) \leqslant m\overline{N}(f^m)$ and $\|\overline{R}(f^m, v)\| \leqslant \|R(f^m, v)\| \leqslant m\|\overline{R}(f^m, v)\|$.

If C is a ring the definition of norm extends to $\mathbb{Z}C$ matrices. For such a matrix M define $\|M\| = \sum_{i,j} \|M_{ij}\|$. Then $\|MN\| \leqslant \|M\|\,\|N\|$ when M and N can be multiplied, and $\|\operatorname{trace} M\| \leqslant \|M\|$.

PROPOSITION 8.1. $N^\infty(f) \leqslant R^\infty(f) < \infty$.

PROOF. The first inequality is clear. For the second,

$$\begin{aligned}
\|R(f^m, v)\| &= \Big\|\sum_k (-1)^k q\big(\operatorname{trace}((\tilde{f}^m)_k)\big)\Big\|_{\mathbb{Z}G_{\phi^m}} \\
&\leqslant \sum_k \|q\big(\operatorname{trace}((\tilde{f}^m)_k)\big)\|_{\mathbb{Z}G_{\phi^m}} \leqslant \sum_k \|\operatorname{trace}((\tilde{f}^m)_k)\|_{\mathbb{Z}G} \\
&\leqslant \sum_k \|(\tilde{f}^m)_k\|_{\mathbb{Z}G} \leqslant \sum_k \|\tilde{f}_k \phi(\tilde{f}_k)\cdots\phi^{m-1}(\tilde{f}_k)\|_{\mathbb{Z}G} \\
&\leqslant \sum_k \|\tilde{f}_k\|^m.
\end{aligned}$$

So

$$R^\infty(f) \leqslant \max_k \operatorname*{growth}_{m\to\infty}\big(\|\tilde{f}_k\|^m\big) < \infty. \qquad \square$$

The next result illustrates how classical complex analysis enters, when the zeta function has a complex rather than a formal variable.

PROPOSITION 8.2 (Lower bound for $R^\infty(f)$). *Let $\rho : G \to \mathrm{GL}_r(\mathbb{C})$ be a representation satisfying $\rho\circ\phi = \rho$. Let $\mu_m \geqslant \sup\{|\operatorname{trace}\rho\Phi(\mathbf{F})| \mid \mathbf{F}$ is an fp class of $f^m\}$, and let $\mu = \operatorname{growth}_{m\to\infty}\mu_m$. For any zero or pole w of the rational function $\zeta(f,\rho) \in \mathbb{C}(t)$, $R^\infty(f) \geqslant \frac{1}{\mu|w|}$. In particular, if ρ is a unitary representation, $R^\infty(f) \geqslant \frac{1}{|w|}$.*

PROOF. $\zeta(f,\rho)(0)=1$, so $w\neq 0$. We have $\log\zeta(f,\rho)(t)=\sum_{m\geqslant 1}\frac{1}{m}R(f^m,\rho)t^m$. Now, growth $L(f^m,\rho)$ is the reciprocal of the radius of convergence of $\log\zeta(f,\rho)$, and is therefore $\geqslant\frac{1}{|w|}$. But $R(f^m,v)=\sum_{\mathbf{F}}\iota(f^m,\mathbf{F})\Phi(\mathbf{F})$, so $|L(f^m,\rho)|\leqslant\sum_{\mathbf{F}}|\iota(f^m,\mathbf{F})|\,|\operatorname{trace}(\rho\Phi(\mathbf{F}))|\leqslant\|R(f^m,v)\|\mu_m$. So $\mu R^\infty(f)\geqslant\frac{1}{|w|}$. □

PROPOSITION 8.3 (Upper bound for $R^\infty(f)$). *Let F_k denote the matrix whose (i,j) entry is $\|(\tilde{f}_k)_{ij}\|_{\mathbb{Z}G}$. Then $R^\infty(f)\leqslant\max_k$(spectral radius of F_k).*

PROOF. As in the proof of Proposition 8.1,

$$\left\|R(f^m,v)\right\|\leqslant\sum_k\left\|\operatorname{trace}\left((\tilde{f}^m)_k\right)\right\|\leqslant\sum_k\operatorname{trace}\left((F_k)^m\right).$$

So $R^\infty(f)\leqslant\max_k(\operatorname{growth}_{m\to\infty}\operatorname{trace}((F_k)^m))=\max_k(\text{spectral radius of }F_k)$.

I will end with a theorem relating asymptotic Nielsen theory to topological entropy.

Let (X,d) be a compact metric space, and let $f:X\to X$ be a map. For $\varepsilon>0$ and $n\geqslant 1$, a subset E of X is *(n,ε)-separated* under f if for each pair $x\neq y$ in E there is $0\leqslant i<n$ such that $d(f^i(x),f^i(y))>\varepsilon$. Let $s_n(\varepsilon,f)$ denote the largest cardinality of (n,ε)-separated subsets E under f. Thus $s_n(\varepsilon,f)$ is the greatest number of orbit segments $\{x,f(x),\dots,f^{n-1}(x)\}$ of length n that can be distinguished one from another when we can only distinguish between points of X that are at least ε apart. Now define

$$h(f,\varepsilon)=\limsup_{n\to\infty}\frac{1}{n}\log s_n(\varepsilon,f)\quad\text{and}\quad h(f)=\lim_{\varepsilon\to 0}h(f,\varepsilon).$$

The number $0\leqslant h(f)\leqslant\infty$, which is easily seen to be independent of the metric d, is the *topological entropy* of f.

If $h(f,\varepsilon)>0$ then, up to resolution $\varepsilon>0$, the number $s_n(\varepsilon,f)$ of distinguishable orbit segments of length n grows exponentially with n. So $h(f)$ measures the growth rate in n of the number of orbit segments of length n with arbitrarily fine resolution.

The following is due to Ivanov [25]. The proof given here is taken from [31]: □

THEOREM 8.4. *Let $f:X\to X$ be a map on a compact polyhedron, and let $h(f)$ denote the topological entropy of f. Then $h(f)\geqslant\log N^\infty(f)$.*

PROOF. Let $\delta>0$ be such that every loop in X of diameter $<2\delta$ is homotopically trivial in X. Let $\varepsilon>0$ be a smaller number such that $d(f(x),f(y))<\delta$ whenever $d(x,y)<2\varepsilon$. Let $E_n\subset X$ be a set consisting of one point from each essential fp class of f^n. Thus $|E_n|=N(f^n)$. By the definition of $h(f)$, it suffices to show that E_n is (n,ε)-separated. Suppose not. Then there would be two fixed points of f^n, $x\neq y\in E_n$, such that $d(f^i(x),f^i(y))\leqslant\varepsilon$ for $0\leqslant i<n$ hence for all $i\geqslant 0$. Pick a path c_i from $f^i(x)$ to $f^i(y)$ of diameter $<2\varepsilon$ for $0\leqslant i<n$ and let $c_n=c_0$. By the choice of δ and ε, $f\circ c_i\simeq c_{i+1}$ for all i, so $f^n\circ c_0\simeq c_n=c_0$. This means x and y are in the same fp class of f^n, contradicting the construction of E_n. □

NOTE. This section is adapted from [31] where a richer discussion with worked examples can be found.

References

[1] H. Bass, *Euler characteristics and characters of discrete groups*, Invent. Math. **35** (1976), 155–196.

[2] R.F. Brown, *The Lefschetz Fixed Point Theorem*, Scott Foresman, Chicago (1971).

[3] R.F. Brown, ed., *Fixed Point Theory and its Applications*, Proceedings of a Conference at Berkeley, California, Contemp. Math., Vol. 72, Amer. Math. Soc., Providence (1988).

[4] M.M. Cohen, *A Course in Simple-Homotopy Theory*, Springer, New York (1973).

[5] D. Dimovski, *One-parameter fixed point indices*, Pacific J. Math. **164** (1994), 263–297.

[6] D. Dimovski and R. Geoghegan, *One-parameter fixed point theory*, Forum Math. **2** (1990), 125–154.

[7] E. Dyer, *Some remarks on fixed point theory*, General Topology and Modern Analysis, Academic Press, New York (1981).

[8] E. Dyer and A.T. Vasquez, *An invariant for finitely generated projectives over* $\mathbb{Z}G$, J. Pure Appl. Algebra **7** (1976), 241–248.

[9] B. Eckmann, *Cyclic homology of groups and the Bass conjecture*, Comment. Math. Helv. **61** (1986), 193–202.

[10] E. Fadell and G. Fournier, eds, *Fixed Point Theory*, Proceedings of a Conference at Sherbrook, Quebec, Canada, Lecture Notes in Math., Vol. 886, Springer, New York (1981).

[11] M. Feckan, *Nielsen fixed point theory and non-linear equations*, J. Differential Equations **106** (1993), 312–331.

[12] D. Fried, *Homological identities for closed orbits*, Invent. Math. **71** (1983), 419–442.

[13] R. Geoghegan, *Fixed points in finitely dominated compacta: the geometric meaning of a conjecture of H. Bass*, Shape Theory and Geometric Topology, Lecture Notes in Math., Vol. 870, Springer, New York (1981), 6–22.

[14] R. Geoghegan and A. Nicas, *Parametrized Lefschetz–Nielsen fixed point theory and Hochschild homology traces*, Amer. J. Math. **116** (1994), 397–446.

[15] R. Geoghegan and A. Nicas, *Trace and torsion in the theory of flows*, Topology **33** (1994), 683–719.

[16] R. Geoghegan and A. Nicas, *Higher Euler characteristics (I)*, Enseign. Math. **41** (1995), 3–62.

[17] R. Geoghegan and A. Nicas, *A Hochschild homology Euler characteristic for circle actions*, K-Theory **18** (1999), 159–178.

[18] R. Geoghegan, A. Nicas and J. Oprea, *Higher Lefschetz traces and spherical Euler characteristics*, Trans. Amer. Math. Soc. **348** (1996), 2039-2062.

[19] D. Gottlieb, *A Certain subgroup of the fundamental group*, Amer. J. Math. **87** (1965), 840–856.

[20] M. Hirsch, *Differential Topology*, Springer, New York (1976).

[21] B. Hughes and A. Ranicki, *Ends of Complexes*, Cambridge Tracts in Math., Vol. 123, Cambridge University Press, Cambridge (1996).

[22] S.Y. Husseini, *Generalized Lefschetz numbers*, Trans. Amer. Math. Soc. **272** (1982), 247–274.

[23] M. Hutchings and Y.-J. Lee, *Circle-valued Morse theory, Reidemeister torsion, and Seiberg–Witten invariants of three manifolds*, Topology **38** (1999), 861–888.

[24] M. Hutchings and Y.-J. Lee, *Circle-valued Morse theory and Reidemeister torsion*, Geom. Topol. **3** (1999), 369–396.

[25] N.V. Ivanov, *Entropy and the Nielsen numbers*, Soviet Math. Dokl. **26** (1982), 63–66.

[26] J. Jezierski, *One-codimensional Wecken type theorems*, Forum Math. **5** (1993), 421–439.

[27] B.J. Jiang, *On the least number of fixed points*, Amer. J. Math. **102** (1980), 749–763.

[28] B.J. Jiang, *Lectures on Nielsen Fixed Point Theory*, Contemp. Math., Vol. 14, Amer. Math. Soc., Providence, RI (1983).

[29] B.J. Jiang, *Fixed points and braids I*, Invent. Math. **75** (1984), 69–74.

[30] B.J. Jiang, ed., *Topological Fixed Point Theory and Applications*, Proceedings of a Conference at Tianjin, China, Lecture Notes in Math., Vol. 1411, Springer, New York (1989).

[31] B.J. Jiang, *Estimation of the number of periodic orbits*, Pacific J. Math. 172 (1996), 151–185.

[32] M.R. Kelly, *Minimizing the number of fixed points for self-maps of compact surfaces*, Pacific J. Math. **126** (1987), 81–123.

[33] T. Kiang, *Theory of Fixed Point Classes*, Springer, Berlin (1989).

[34] P.A. Linnell, *Decomposition of augmentation ideals and relation modules*, Proc. London Math. Soc. (3) **47** (1983), 83–127.

[35] C.K. McCord, ed., *Nielsen Theory and Dynamical Systems*, Contemp. Math., Vol. 152, Amer. Math. Soc., Providence (1993).

[36] J. Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia, Charlottsville (1965).

[37] J. Milnor, *Infinite cycle coverings*, Conference on the Topology of Manifolds (Michigan State Univ., E. Lansing, Mich., 1967), Prindle, Weber & Schmidt, Boston (1968), 115–133.

[38] J. Nielsen, *Untersuchungen zur Topologie des geschlossen zweiseitigen Flache*, I, II, III, Acta Math. **50** (1927), 189–358; **53** (1929), 1–76; **58** (1932), 87–167.

[39] A. Pajitnov, *Simple homotopy type of the Novikov complex and Lefschetz zeta-functions of the gradient flow*, Russian Math. Surveys **54** (1999), 119–169.

[40] A. Pajitnov, *C^0-generic properties of boundary operators in the Novikov complex, pseudoperiodic topology*, Amer. Math. Soc. Transl. Ser. 2 **197** (1999), 29–115.

[41] A. Pajitnov, *Closed orbits of gradient flows and logarithms of non-Abelian Witt vectors*, to appear in K-Theory.

[42] A. Pajitnov and A. Ranicki, *The Whitehead group of the Novikov ring*, to appear in K-Theory.

[43] K. Reidemeister, *Automorphismen von Homotopiekettenringen*, Math. Ann. **112** (1936), 586–593.

[44] J.A. Schaeffer, *Relative cyclic homology and the Bass conjecture*, Comment. Math. Helv. **67** (1992), 214–225.

[45] D. Schuetz, *Gradient flows of closed 1-forms and their closed orbits*, to appear in Forum Math.

[46] D. Schuetz, *One-parameter fixed point theory gradient flows of closed one-forms*, Preprint.

[47] S.P. Singh, ed., *Nonlinear Functional Analysis and its Applications*, NATO-ASI Series C, Vol. 173, Kluwer Academic Publishers (1986).

[48] E. Spanier, *Algebraic Topology*, McGraw-Hill, New York (1966).

[49] J. Stallings, *Centerless groups – an algebraic formulation of Gottlieb's theorem*, Topology **4** (1965), 129–134.

[50] F. Wecken, *Fixpunktklassen*, I, Math. Ann. **117** (1941), 659–671; II, Math. Ann. **118** (1942), 216–234; III, Math. Ann. **118** (1942), 544–577.

[51] J.E. West, *Mapping Hilbert cube manifolds to ANR's: a solution to a conjecture of Borsuk*, Ann. of Math. **106** (1977), 1–18.

[52] X. Zhang, *The least number of fixed points can be arbitrarily larger than the Nielsen number*, Acta Sci. Natur. Univ. Pekinensis (1986), 15–25.

CHAPTER 12

Mapping Class Groups*

Nikolai V. Ivanov

Michigan State University, Department of Mathematics, Wells Hall, East Lansing, MI 48824-1027, USA
E-mail: ivanov@math.msu.edu

Contents

---

*Supported in part by the NSF Grants DMS-9401284 and DMS-9704817.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

The *mapping class group* Mod_S of an orientable surface S is defined as the group of isotopy classes of orientation-preserving diffeomorphisms $S \to S$. In addition to being a central object of the topology of surfaces (cf. 2.1), these groups also play an important role in the theory of Teichmüller spaces and in algebraic geometry, where they are known under the name *Teichmüller modular groups* or simply *modular groups*. Our notations are derived from the latter terminology. There are several closely related groups, which also deserve the name of the mapping class groups (or Teichmüller modular groups). First of all, one may consider the *extended mapping class group* $\mathrm{Mod}_S^{\diamond}$ of S defined as the group of the isotopy classes of *all* diffeomorphisms $S \to S$. The *pure mapping class group* PMod_S of S is defined as the group of isotopy classes of all orientation-preserving diffeomorphisms $S \to S$ preserving setwise all boundary components of S. Finally, one may consider the group $\mathcal{M}_S$ of all (orientation-preserving) diffeomorphisms $S \to S$ fixed on the boundary ∂S, considered up to isotopies fixed on the boundary. If $\partial S \neq \emptyset$, then diffeomorphisms fixed on ∂S are automatically orientation-preserving. If $\partial S = \emptyset$, then, of course, $\mathcal{M}_S = \mathrm{Mod}_S$. All these groups could be also defined as the 0-th homotopy groups of suitable diffeomorphisms groups of S. For example, $\mathrm{Mod}_S = \pi_0(\mathrm{Diff}(S))$, where $\mathrm{Diff}(S)$ is the group of all orientation-preserving diffeomorphisms of S considered with, for example, C^{∞}-topology (or any other reasonable topology).

The study of the mapping class groups was initiated in the 1920-ies by Dehn [43–45] and Nielsen [191–193]. Although their work had some common themes (cf., for example, the Dehn–Nielsen–Baer theorems in 2.9), in general their approaches were fairly different. Dehn was interested in the properties of the mapping class group as a whole, addressing, for example, such questions as the existence of a finite set of generators. He developed and exploited an important tool for this purpose: the action of the mapping class group Mod_S on the collection of the isotopy classes of all circles on S. He called this collection the *arithmetic field* of S. On the contrary, Nielsen was mainly interested in understanding the fine structure of the individual elements of Mod_S. His methods draw heavily on hyperbolic geometry (a tool also favored by Dehn). For quite a while, the work of both Dehn and Nielsen was apparently forgotten. The ideas of Dehn related to the arithmetic field of a surface found a natural continuation in the ideas of Harvey [86,87] about the *complex of curves* of a surface, which is nothing else but the arithmetic field made in a simplicial complex in a natural way (see 3.1). A closely related object was considered in an influential paper of Hatcher and Thurston [91]. The ideas of Nielsen were partially rediscovered, extended and brought to an essentially complete form by Thurston in his theory of surface diffeomorphisms [53,222] (see Section 7). Later on, his theory was applied also to the structure of the mapping class group (and not only to its individual elements).

In the present notes we have adopted a point of view going back to that of Dehn. Compared to the other thread, going back to Nielsen, it is more elementary and allows us to reach more quickly some really deep results about mapping class groups. Everyone interested in the mapping class groups eventually should learn the other point of view also (especially Thurston's theory of surface diffeomorphisms). We hardly do justice to it in a chapter devoted to Thurston's theory and some of its applications. If one takes Dehn's terminology seriously, our exposition is slanted toward the *arithmetic* of our subject, which seems to be only natural for an introduction. The analogy with the *arithmetic groups* (see 9.1 for a definition), briefly touched on in Section 9, fits nicely in this approach (for example, the main results of Section 6 are, in fact, analogues of some fundamental theorems about arithmetic groups), giving a deeper meaning to the word *arithmetic* in this context.

Our exposition is centered around three topics: the Dehn–Lickorish theorem providing a finite set of generators of Mod_S for closed S; Harer's theorem computing the so-called virtual cohomology dimension of Mod_S; and the author's theorem describing all automorphisms of Harvey's complex of curves. The first two topics are presented with complete proofs and a detailed review of the prerequisites; for the third topic only an outline is given. The proofs we give for the theorems of Dehn–Lickorish and Harer contain some new ideas and may be interesting even for experts. The table of contents gives a good idea of the topics included in these notes. The limitations of time and space prevented us from giving a more detailed treatment of many included topics and from discussing other topics, equally or, may be, even more important. As we already mentioned, our treatment of Thurston's theory is hardly adequate (although we hope that it is still useful); the same applies to the aspects of the theory of the Teichmüller spaces related to the mapping class groups. Nothing is said about relations with algebraic geometry and mathematical physics; we refer to the recent survey of Hain and Looijenga [72] for this. Two omitted topics are very closely related to the ones included. In order not to pass them in complete silence, we very briefly discuss them here.

The first one is the study of the so-called *Torelli group*, the subgroup of Mod_S consisting of the isotopy classes of diffeomorphisms of S acting trivially on $H_1(S, \mathbf{Z})$. The pioneering work in this field is due to Johnson. In [115] he proved that the Torelli group of S is finitely generated if the genus of S is at least 3 and S has no more than 1 boundary component, and in [116,117] he computed the first homology group of the Torelli group (under the same restrictions). See [114] for a survey of this work. It was continued, in particular, in the algebro-geometric work of Hain; see [70], and especially the remarkable paper [71], where Hain proved the existence of a finite presentation for a Lie algebraic analogue of the Torelli group. Note that the question about the existence of a finite presentation of the Torelli group itself remains open.

The second topic is the theory and applications of the characteristic classes of surface bundles (which are nothing other than the elements of the cohomology group $H^*(\mathrm{Mod}_S)$), constructed independently by Mumford [188], Morita [168,169] and Miller [164]. Morita [168,169] and Miller [164] proved the nontriviality of these classes, and Morita [168,169] found a spectacular application of this nontriviality: he proved that the canonical homomorphism $\mathrm{Diff}(S) \to \mathrm{Mod}_S$ admits no section. In his further papers (see, for example, [170–173,177–179]) Morita studied these characteristic classes and applied them to the theory of the Torelli group (extending in some respects the above mentioned work of Johnson) and to the Casson invariant of homology 3-spheres and related problems of the topology of 3-manifolds. We refer to his own surveys [174–176] for a discussion of these results.

2. Topology of surfaces

In this section we attempt to outline the material which should be at the foundation of the theory of the mapping class groups. All this material is fairly classical, although some results (for example, the ones about spaces of diffeomorphisms and embeddings) are not so widely known as they deserve. Our discussion is more broad than is needed for later applications, in order to get closer to a coherent picture; still, our treatment is far from being complete. The proofs in this section are not to be taken too seriously; mostly, they are only outlines or just indications of the ideas involved in the real proofs.

2.1. *The dimension 2*

In order to put the topology of surfaces in a perspective, let us start with a particular view of the topology of manifolds in general.

The central problem of the topology of manifolds is the problem of classification of manifolds. It is usually considered to be more of less solved for at least simply connected manifolds of dimension $\geqslant 5$. Another major problem is the problem of isotopy classification of diffeomorphisms. Consideration of diffeomorphisms up to isotopy distinguishes topology from other subjects, especially from the theory of smooth dynamical systems, which concentrates on the classification of diffeomorphisms up to conjugacy. A satisfactory solution of the isotopy classification problem for diffeomorphisms is known only for some special classes of manifolds. The problem which naturally comes after the classification of diffeomorphisms is the problem of understanding of the homotopy type of the diffeomorphism group of a manifold. It is extremely difficult even for such manifolds as spheres. These problems actually form a natural series. Namely, the classification of manifolds, the isotopy classification of diffeomorphisms, and the computation of homotopy groups $\pi_i(\mathrm{Diff}(M))$ for $i > 0$, where $\mathrm{Diff}(M)$ is the group of all diffeomorphisms $M \to M$ (with, say, C^∞-topology), form a series of problems, parallel and related to the study of functors K_1, K_2, and K_i for $i > 2$ of the algebraic K-theory. We can add also to this row the problem of the existence of a smooth manifold within a given homotopy type and the functor K_0.

In the dimension 2, the solution of the classification problem is well known and is, in fact, a part of any introductory course in topology. The homotopy groups $\pi_i(\mathrm{Diff}(M))$ with $i > 0$ are known for all compact two-dimensional manifolds, and, moreover, most of them are equal to 0. We will discuss these results in 2.7. This leaves us in the dimension 2 with only one of the above problems – the problem of the classification of diffeomorphisms. Interpreted broadly, this problem naturally includes a search for an understanding of the mapping class groups $\pi_0(\mathrm{Diff}(M))$, which is the main focus of our survey. But, before we concentrate on $\pi_0(\mathrm{Diff}(M))$, we will review some other parts of the topology of surfaces, partly because they are needed to understand $\pi_0(\mathrm{Diff}(M))$ and partly for their own sake.

2.2. *Classification of surfaces*

Throughout the paper, all surfaces are assumed to be orientable and, unless the contrary is obvious from the context, compact.

There is no doubt that the reader is familiar with the classification of surfaces. Let us recall, for the sake of completeness, that an orientable connected surface S is determined up to a diffeomorphism by its genus g and the number of the boundary components b. Following Riemann, the genus can be defined as the maximal number of disjoint circles on S such that these circles do not separate S, i.e., the complement in S of the union of these circles is connected. The genus g and the number of boundary components b are related to the Euler characteristic $\chi(S)$ of S by the well known formula $\chi(S) = 2-2g-b$. A surface of genus g with b boundary components is also known as a *sphere with g handles and b holes;* the terminology is justified by a well known picture.

The standard proof of this classification theorem, based on triangulations and cutting and pasting arguments, provides a cannonical model of a surface of genus g with b boundary components. Namely, such a surface can be obtained from a $4g$-gon by a well known identification of sides and removing b disjoint discs from the interior. The identification of sides is described by the word $a_1b_1a_1^{-1}b_1^{-1}\dots a_gb_ga_g^{-1}b_g^{-1}$ and this leads to a presentation of $\pi_1(S)$:

$$\begin{aligned}\pi_1(S) = \langle &a_1, b_1, \dots, a_g, b_g, d_1, \dots, d_b \mid \\ &a_1b_1a_1^{-1}b_1^{-1}\dots a_gb_ga_g^{-1}b_g^{-1}d_1\dots d_b = 1\rangle.\end{aligned}$$

If $b > 0$, then one can eliminate one d_i from this presentation and conclude that $\pi_1(S)$ is a free group. One can obtain the same result in a more topological manner, by observing (using the above $4g$-gon model, for example) that S has a 1-dimensional CW-complex as a deformation retract. But considering $\pi_1(S)$ for surfaces with boundary as just a free group is to a great extent misleading, because in this way we lose all information about the boundary. It is much better to keep the track of the boundary by introducing the so-called *peripheral structure* – the set of $2b$ conjugacy classes of elements of $\pi_1(S)$ corresponding to loops going once round a boundary component. It is also useful to consider the *oriented peripheral structure* – the set of b conjugacy classes of elements of $\pi_1(S)$ corresponding to loops going once round a boundary component in the direction prescribed by a fixed orientation of S (which induces an orientation of the boundary ∂S).

Note that all generators in the above presentation can be obviously represented by *embedded* loops, i.e., by loops $f:[0,1] \to S$ such that $f(x) = f(y)$ only if $x = y$ or $\{x, y\} = \{0, 1\}$ (of course $f(0) = f(1)$ is the base point).

2.3. *Circles and arcs on surfaces*

Circles and arcs on surfaces and the action of diffeomorphisms on them are some of the main tools (and objects) of the topology of surfaces. By *circles* we understand the *embedded* circles, i.e. submanifolds diffeomorphic to the standard circle. Similarly, by *arcs* we understand submanifolds diffeomorphic to the interval $[0, 1]$. Usually, one restricts attention to properly embedded arcs; an arc I on a surface S is called *properly embedded* (or simply *proper*) if $\partial I = I \cap \partial S$ and I is transverse to ∂S.

It is convenient to single out circles and arcs considered to be trivial. We call a circle C on a surface S *trivial* if C is either contractible in S, or is homotopic to some boundary component of S. Similarly, we call an arc I *trivial* if it can be deformed into the boundary of S in such a way that its boundary ∂I stays in ∂S. Also, it is important to distinguish between separating and nonseparating circles. We call a circle C on S *nonseparating* if $S \setminus C$ is connected, and *separating* otherwise.

As we noticed in 2.2 all generators in the standard presentation of $\pi_1(S)$ can be represented by embedded loops. The image of an embedded loop is a circle, and we call an embedded loop *separating* or *nonseparating* if this image is separating, or, respectively, nonseparating. The obvious loops representing $a_1, b_1, \ldots, a_g, b_g$ are nonseparating, while the obvious loops representing $d_1, \ldots, d_b$ are separating (in fact, their images can be deformed into the corresponding boundary components). But one can easily represent elements $a_1 d_1, \ldots, a_1 d_b$ by embedded nonseparating loops. It follows that $\pi_1(S)$ can be generated by elements represented by nonseparating embedded loops. This remark will be used in our proof of the Dehn–Lickorish theorem (see the proof of Theorem 4.2.C).

If C is a circle on S or, more generally, union of several disjoint circles, we can cut S along C and get a new surface, which will be denoted by S_C. If C is a nonseparating circle, then S_C is connected. If C is a separating circle, then S_C consists of two components. Note that we can always reconstruct S from S_C by gluing along the boundary components of S_C resulting from C. More generally one can cut S also along proper arcs and, more generally, disjoint unions of circles and proper arcs. If (proper) arcs are involved, the result of cutting is a smooth surface with corners.

Suppose now that C is a nonseparating circle. Then S_C is a connected surface having two more boundary components than S. The additivity of the Euler characteristic implies that S_C has the same Euler characteristic as S. Hence, by the classification of surfaces, S_C is determined up to a diffeomorphism by S (and the fact that C is nonseparating).

LEMMA 2.3.A. *If C, C' are two nonseparating circles on S, then there is a diffeomorphism $f : S \to S$ such that $f(C) = C'$.*

PROOF. By the remarks preceding the lemma, S_C and $S_{C'}$ are diffeomorphic. Let us fix an orientation of S; it induces an orientation of both S_C and $S_{C'}$. Since every orientable surface obviously admits an orientation-reversing diffeomorphism, we can always choose an orientation-preserving diffeomorphism $F : S_C \to S_{C'}$. Let C_- and C_+ be two boundary components of S_C resulting from C, and let $C'_- = F(C_-)$, $C'_+ = F(C_+)$. Let $I : C_- \to C_+$ and $I' : C'_- \to C'_+$ be two gluing diffeomorphisms giving S back from S_C and $S_{C'}$, respectively. If F agrees with these gluings, i.e., if $F \circ I = I' \circ F$ on C_-, then F gives rise (by gluing) to a diffeomorphism $f : S \to S$ such that $f(C) = C'$. Note that both I and I' are orientation-reversing (assuming that C_-, C_+, C'_-, C'_+ are oriented as boundary components of S_C, $S_{C'}$) and F is orientation-preserving on both C_- and C_+. It follows that diffeomorphisms $F \circ I$, $I' \circ F : C_- \to C'_+$ are both orientation-reversing. Since every two orientation-reversing (as also every two orientation-preserving) diffeomorphisms of a circle are isotopic, we can find a diffeomorphism $F' : C_+ \to C'_+$ isotopic to $F \mid C_+$ and such that $F' \circ I = I \circ F \mid C_-$. By extending the isotopy between $F \mid C_+$ and F' to an isotopy of F (and without changing $F \mid C_-$), we get a diffeomorphism $F : S_C \to S_{C'}$ which agrees with gluings. As we noted above, this is sufficient to complete the proof. □

If C is nonseparating, then there is a circle C' transversely intersecting C at exactly one point. (To construct such a C', take a small arc I intersecting C transversely at one point in the interior of I and connect the two ends of I by an arc in $S \setminus C$.) In particular, the homological intersection number $C \cdot C'$ is ± 1. This immediately implies that C is nontrivial.

LEMMA 2.3.B. *A circle C contractible in S bounds a disc in S. A circle C homotopic to a boundary component B of S, bounds, together with B, an annulus in S. In particular, a circle is non-trivial if and only if it does not bound a disc in S and does not bound an annulus in S together with a boundary component.*

PROOF. First, notice that a trivial circle C is always separating by the remarks preceding the lemma. For such a C, we can represent S as the union of two surfaces S_1, S_2 intersecting only along their common boundary component C. If C is contractible in S, then van Kampen's theorem (together with the basic properties of the amalgamated products) implies that C is contractible in one of the surfaces S_1, S_2, say in S_1. In view of 2.2, any embedded loop with image C is freely homotopic to a loop representing d_i or d_i^{-1} for some generator d_i of $\pi_1(S_1)$. Because C is contractible in S_1, this implies that $d_i = 1$. Obviously, this is possible only if the genus of S_1 is 0 and S_1 has only one boundary component (namely, C), i.e., only if S_1 is a disc. In this case C bounds a disc in S, namely S_1. A similar argument applies if C is homotopic to a boundary component. We leave it to the reader. □

2.4. *Pants decompositions*

With few exceptions, a surface has negative Euler characteristic (by the classification of surfaces). The simplest such surface is a disc with two holes, often called *a pair of pants*, or simply *pants*; see Figure 1(a). As the next theorem shows, pairs of pants are the elementary building blocks of all surfaces of negative Euler characteristic.

THEOREM 2.4.A. *If S is a surface of negative Euler characteristic (i.e., if S is not a sphere, a torus, a disc, or an annulus), then there is a collection of disjoint circles on S such that the result S_C of cutting S along their union C is a disjoint union of pairs of pants. In other words, C decomposes S into pairs of pants.*

PROOF. Using the classification of surfaces, one can easily exhibit the required collection of circles in each case. Let us outline a proof independent of the classification. Choose a

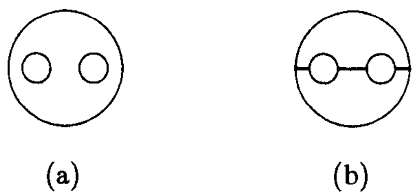


Fig. 1.

Morse function $f: S \to \mathbf{R}$ on S such that f is equal to 0 on ∂S, f is nonnegative and the values of f at different critical points are different. Let $0 = a_0 < a_1, \ldots < a_n$ be a sequence of noncritical values of f such that every interval $[a_{i-1}, a_i]$, $1 \leqslant i \leqslant n$, contains exactly one critical value of f and $\operatorname{Im} f \subset [0, a_n)$. Consider a component of the preimage $f^{-1}([a_{i-1}, a_i])$. If it does not contain a critical point of f, it is an annulus. If it contains a critical point of f of index 0 or 2, then it is a disc. Finally, if it contains a critical point of f of index 1, then it is easily seen to be a disc with two holes. So, if we cut S along the union of the components of all preimages $f^{-1}(a_i)$, $i \geqslant 1$, the resulting surface will consist of discs, annuli, and pairs of pants. We can simplify this decomposition of S into discs, annuli, and pairs of pants in the following manner. Note that gluing a disc to a boundary component of a disc, annulus or a pair of pants results in a sphere, a disc, or an annulus respectively. Also, gluing an annulus to a boundary component of any surface does not change it up to a diffeomorphism. Hence, by repeatedly gluing discs and annuli to other pieces of the decomposition, we will eventually reach a decomposition of S either into several pairs of pants, or into two discs glued together (in this case S is a sphere), or into an annulus glued to itself (in this case S is a torus), or into just one disc or annulus. To complete the proof, note that a gluing operation corresponds to removing a circle from our collection of cutting circles. □

This theorem can be used to give an alternative proof of the classification of surfaces. Let us cut S along the union of several circles into a finite collection of pairs of pants, as in the theorem. Then let us reassemble S from this collection in the following manner. Start with some pair of pants and glue to it the other pairs of pants, one at time, and only along one boundary component at a time, as long as this is possible. At every stage we will have, as one can easily see by induction, a disc with several holes. When no further such gluing is possible, the remaining gluings identify some pairs of boundary components of our disc with holes. If there are g such pairs and b boundary components not in any such pair, then the result of the gluing of these pairs is, obviously, a sphere with g handles and b holes, i.e., a surface of genus g with b boundary components.

Sometimes it is useful to decompose pairs of pants further. Namely, three nontrivial arcs, shown in Figure 1(b), decompose a pair of pants into two "topological hexagons" (which are smooth manifolds with corners and are homeomorphic, but not diffeomorphic to a disc). Combining this decomposition with Theorem 2.4.A, we can decompose S into such topological hexagons.

From the perspective of 3-dimensional topology, nontrivial circles and arcs are analogues of incompressible surfaces, and the above decomposition of S into hexagons is an analogue of the Haken decomposition of a Haken manifold into topological balls. In fact, since pairs of pants are simple enough, one rarely needs to use nontrivial arcs to decompose a surface into simpler pieces. Note that from the point of view of this analogy all surfaces, with the exception of discs and spheres, are Haken.

2.5. *Geometric structures on surfaces*

By a *geometric structure* on a surface S we understand simply a Riemannian metric on S having constant curvature and geodesic boundary. By scaling the metric, if necessary, we can assume that its curvature is equal to 1, 0 or -1. It turns out that every surface admits a geometric structure. For surfaces of nonnegative Euler characteristic this is very simple. For a sphere, it is sufficient to take the metric of the standard unit sphere in $\mathbf{R}^3$. For a disc, it is sufficient to take the metric of a hemisphere of a standard sphere. For an annulus, one may take a product metric on $S^1 \times [0, 1]$. For a torus, one may take the metric on $\mathbf{R}^2/\mathbf{Z}^2$ induced from the standard Euclidean metric on $\mathbf{R}^2$. In particular, spheres and discs admit metrics of (constant) positive curvature, and annuli and tori admit metrics of curvature 0. The really interesting case is the case of surfaces of negative Euler characteristic. They all admit *hyperbolic structures*, i.e., geometric structures with metric of curvature -1; see Theorem 2.5.B below.

Geometric structures on surfaces can be used on two levels. On the first level, the very existence of a geometric structure is useful. For example, fixing a geometric structure on a surface S allows us to represent isotopy classes of circles by geodesics; for hyperbolic structures such representatives are unique. Moreover, the following lemma holds.

LEMMA 2.5.A. *Suppose that S is endowed with a geometric structure. Then every nontrivial circle on S is isotopic to a geodesic circle on S. If S is endowed with a hyperbolic structure, then such a geodesic circle is unique. If two nontrivial circles are disjoint, then any two geodesic circles isotopic to them are either disjoint or equal.*

We refer to [53, Exposé 3, §§I and III] for a proof. On a second level, one considers all possible geometric structures on S simultaneously. It is convenient to identify geometric structures on S which differ one from another in a trivial way; namely, to identify two geometric structures if one is the pull-back of the other by a diffeomorphism $S \to S$ isotopic to the identity. The resulting set of equivalence classes is the *Teichmüller space of S*. We explore this point of view in Section 5.

THEOREM 2.5.B. *If a surface has negative Euler characteristic, then it admits a hyperbolic structure.*

PROOF. There are (at least) two ways to prove this result. The first one works best for closed surfaces S. Represent S as a result of the standard identification of sides of a $4g$-gon; cf. 2.2. One can realize this $4g$-gon as a regular polygon on the hyperbolic plane with all angles equal to $2\pi/4g = \pi/2g$. Note that the hyperbolic plane has a canonical Riemannian metric of constant negative curvature (in the upper half plane model, it is the metric $(\mathrm{d}x^2 + \mathrm{d}y^2)/y^2$ on $\{(x, y) \in \mathbf{R}^2\colon\ y > 0\}$); this metric induces a Riemannian metric on S because the lengths of glued edges are equal and the sum of angles is equal to 2π (notice that all vertices of our $4g$-gon are identified in the gluing process). Clearly, this metric on S has constant negative curvature. Note that it is not necessary to consider a regular $4g$-gon; a weaker condition would suffice. For the details of this construction we refer to [217, Sections 5.5 and 5.6].

Now, one can use Lemma 2.5.A to deal with the surfaces with boundary. Given a surface with boundary S, we can endow the double $\mathrm{d}S$ of S with a hyperbolic structure. Recall that $\mathrm{d}S$ is the result of gluing S with a copy of itself along the boundary; in particular, $\mathrm{d}S \supset S$. Lemma 2.5.A implies that ∂S is isotopic in $\mathrm{d}S$ to a union of disjoint geodesic circles. If we cut dS along this union, we get two surfaces diffeomorphic to S, each of them endowed with a hyperbolic structure.

Another construction works equally well both for closed surfaces and for surfaces with boundary and is based on Theorem 2.4.A. Let us prove first that a pair of pants P admits a hyperbolic structure such that all three boundary components have the same length. To this end, we represent P as the union of two topological hexagons as in Figure 1(b). Let us realize these hexagons as two (isometric) regular hexagons in the hyperbolic plane with all angles equal to $\pi/2$. Then gluing P from these hexagons endows P with a hyperbolic structure (one should take isometries as the gluing maps; this is possible because lengths of edges are all equal); the boundary is geodesic because angles of hexagons are equal to $\pi/2$. Clearly, the lengths of all boundary components of P are equal to twice the length of an edge of our hexagons; in particular, the lengths of all boundary components of P are equal. Now, given a surface S of negative Euler characteristic, we can decompose S into pairs of pants as in Theorem 2.4.A. Let us endow all pairs of pants of this decomposition with the hyperbolic structure we just constructed. Since the lengths of all boundary components of these pairs of pants are equal, one can take isometries as gluing maps and then these hyperbolic structures on pieces of S induce a hyperbolic structure on S itself. For the details, and a more general version of this approach, we refer to [29, Section 1.7 and Chapter 3]. It can be easily generalized to give a description of all hyperbolic structures on S; see [29, Chapters 3 and 6], and [53, Exposé 7 and Section 5]. □

2.6. *Spaces of diffeomorphisms and embeddings: generalities*

For an orientable manifold M we denote by $\mathrm{Diff}^\diamond(M)$ the group of all diffeomorphisms $M \to M$, and by $\mathrm{Diff}(M)$ the subgroup of all orientation-preserving diffeomorphisms. (Since we are interested mainly in the groups $\mathrm{Diff}(M)$, we chose for them a simpler notation.) For $X \subset M$ we denote by $\mathrm{Diff}^\diamond(M \operatorname{fix} X)$ the subgroup of diffeomorphisms $f : M \to M$ fixed on X, i.e., such that $f(x) = x$ for all $x \in X$. We denote by $\mathrm{Diff}(M \operatorname{fix} X)$ the intersection $\mathrm{Diff}^\diamond(M \operatorname{fix} X) \cap \mathrm{Diff}(M)$. The most important case is that of $X = \partial M$; we denote the groups $\mathrm{Diff}^\diamond(M \operatorname{fix} \partial M)$ and $\mathrm{Diff}(M \operatorname{fix} \partial M)$ by $\mathrm{Diff}^\diamond(M \operatorname{fix} \partial)$ and $\mathrm{Diff}(M \operatorname{fix} \partial)$ respectively.

If N is a submanifold of M, we denote by $\mathrm{Emb}^\diamond(N, M)$ the set of all embeddings $N \to M$ which can be extended to a diffeomorphism $M \to M$. If N has codimension 0 in M (the main case is that of a disc in the interior of M), then it makes sense to speak about orientation-preserving embeddings $N \to M$, and we denote by $\mathrm{Emb}(N, M)$ the subset of all orientation-preserving embeddings. By $\mathrm{Sub}(N, M)$ we denote the set of submanifolds of M diffeomorphic to N by a diffeomorphism which extends to M (i.e., the set of all images of embeddings from $\mathrm{Emb}^\diamond(N, M)$).

As usual, all diffeomorphisms and embeddings are assumed to be of the class C^∞, and the above groups of diffeomorphisms and sets of embeddings are considered together with their C^∞-topology. $\mathrm{Sub}(M, N)$ is considered with the topology of a quotient space of $\mathrm{Emb}^\diamond(N, M)$. As is well known, groups of diffeomorphisms with the C^∞-topology are topological groups. Also, for a submanifold N of M the natural map $\mathrm{Diff}^\diamond(M) \to \mathrm{Emb}^\diamond(N, M)$ given by the restriction of diffeomorphisms to N is continuous.

It is easy to see that if M is a surface, then $\mathrm{Mod}_S = \pi_0(\mathrm{Diff}(S))$, $\mathrm{Mod}_S^\diamond = \pi_0(\mathrm{Diff}^\diamond(S))$ and $\mathcal{M}_S = \pi_0(\mathrm{Diff}(S \operatorname{fix} \partial))$.

THEOREM 2.6.A. *Let N be submanifold of M such that ∂N intersects ∂M along several components of ∂N and N is transverse to ∂M. Then the natural map $\mathrm{Diff}^\diamond(M) \to \mathrm{Emb}^\diamond(N, M)$ is a Serre fibration. If N is of codimension 0 in M, then the natural map $\mathrm{Diff}(M) \to \mathrm{Emb}(N, M)$ is a Serre fibration. If N is of codimension 0 in M and is contained in the interior of M, then the natural map $\mathrm{Diff}(M \operatorname{fix} \partial) \to \mathrm{Emb}(N, M)$ is also a Serre fibration.*

PROOF. Recall that a continuous map is a Serre fibration if the homotopy lifting property holds for cubes. For $\mathrm{Diff}^\diamond(M) \to \mathrm{Emb}^\diamond(N, M)$ the homotopy lifting property for 0-dimensional cubes amounts to the extension of isotopies of N in M to isotopies of the whole M. This extension property is well known and easy to prove; see for example, [128, Theorem II.5.2]. The homotopy lifting property for higher-dimensional cubes is essentially a multiparameter version of this isotopy extension property and can be proved by a direct generalization of the standard proof of the usual isotopy extension property. Also, the second statement of the theorem obviously follows from the first. The last statement of the theorem can be proved in exactly the same manner as the first one. □

THEOREM 2.6.B. *The natural map $\mathrm{Emb}^\diamond(N, M) \to \mathrm{Sub}(N, M)$, $\mathrm{Emb}(N, M) \to \mathrm{Sub}(N, M)$, given by the formula $f \mapsto f(N)$, are Serre fibrations.*

This theorem can be proved using a variant of the ideas of the proof of Theorem 2.6.A.

The fibers of the fibrations $\mathrm{Diff}^\diamond(M) \to \mathrm{Emb}^\diamond(N, M)$ and $\mathrm{Diff}(M) \to \mathrm{Emb}(N, M)$ are easy to identify. They are $\mathrm{Diff}^\diamond(M \operatorname{fix} N)$ and $\mathrm{Diff}(M \operatorname{fix} N)$ respectively. In addition, if N is a submanifold of codimension 0 contained in the interior of M, then we can identify $\mathrm{Diff}^\diamond(M \operatorname{fix} N)$ and $\mathrm{Diff}(M \operatorname{fix} N)$ with $\mathrm{Diff}^\diamond(M \setminus \operatorname{int} N \operatorname{fix} \partial N)$ and $\mathrm{Diff}(M \setminus \operatorname{int} N \operatorname{fix} \partial N)$ respectively. The fibers of the maps $\mathrm{Emb}^\diamond(N, M) \to \mathrm{Sub}(N, M)$, $\mathrm{Emb}(N, M) \to \mathrm{Sub}(N, M)$ are, obviously, $\mathrm{Diff}^\diamond(N)$ and $\mathrm{Diff}(N)$ respectively.

Now, let us fix a point $m \in M$ and consider the space FM of all orientation-preserving isomorphisms $T_m M \to T_x M$, $x \in M$ (where $T_x M$ is, as usual, the tangent space to M at x). There is an obvious projection $FM \to M$; moreover, FM is a principal $\mathrm{GL}^+(n, \mathbf{R})$-bundle over M, where $n = \dim M$. Let $d : \mathrm{Diff}(M) \to FM$ be the map assigning to a diffeomorphism $M \to M$ its differential at m. If N is a submanifold of codimension 0 and $m \in N$, then there is a similar map $d : \mathrm{Emb}(N, M) \to FM$.

THEOREM 2.6.C. *The maps $d : \mathrm{Diff}(M) \to FM$ and $d : \mathrm{Emb}(N, M) \to FM$ are Serre fibrations. If N is a disc of codimension 0 in the interior of M, then the second map is a weak homotopy equivalence.*

PROOF. We may consider points of FM as framed points in M. Then the homotopy lifting property for d becomes a version of the (multiparameter) isotopy extension property for isotopies of points, which is not much more difficult to prove than the usual isotopy extension property. The second statement of the theorem is essentially a multiparameter version of a well known theorem to the effect that there is exactly one isotopy class of orientation-preserving embeddings of a codimension 0 disc in an orientable manifold. See, for example, [128, Corollary III.3.6]. Again, it is easy to add parameters to the usual proof. □

If $\dim M = 2$, it convenient to fix a Riemannian metric on M and consider the unit tangent bundle $UT(M)$ instead of FM. Let us fix a vector $v \in T_m M$. Then the evaluation at v followed by the normalization defines a map $FM \to UT(M)$, which is a homotopy equivalence (if $\dim M = 2$). Also, define $u : \mathrm{Diff}(M) \to UT(M)$ as the map assigning to a diffeomorphism $M \to M$ the normalized image of v under its differential at m. Similarly define $u : \mathrm{Emb}(N, M) \to UT(M)$.

THEOREM 2.6.D. *The maps* $u : \mathrm{Diff}(M) \to UT(M)$ *and* $u : \mathrm{Emb}(N, M) \to UT(M)$ *are Serre fibrations. If* N *is a disc of codimension* 0 *in the interior of* M *(and* $\dim M = 2$*), then the second map is a weak homotopy equivalence.*

PROOF. The first statement is similar to the first statement of Theorem 2.6.C and, in fact, holds in all dimensions. The second statement follows from the second statement of Theorem 2.6.C and the fact that the above map $FM \to UT(M)$ is a homotopy equivalence. □

The results of this section go back to Cerf [31] and, for closed manifolds, Palais [196]. In fact, they proved that our Serre fibrations are actually fiber bundles. But we need only the Serre fibration property, because we are interested only in the homotopy groups of spaces of diffeomorphisms and embeddings (in fact, only in π_0 and π_1), and this property is much easier to prove.

2.7. *Spaces of diffeomorphisms and embeddings: surfaces*

In this section, we describe the weak homotopy type of some spaces of diffeomorphisms of surfaces and embeddings of circles. While all these results are known to be true for the usual homotopy type instead of the weak one, the results about the weak homotopy type are easier to prove (in particular, there is no need to know that Serre fibrations from 2.6 are fiber bundles), and only they are needed for the applications we have in mind.

We start with two preliminary results about 1-dimensional manifolds.

LEMMA 2.7.A. $\mathrm{Diff}(D^1 \operatorname{fix} \partial)$ *is contractible, where* D^1 *is the* 1*-dimensional disc.*

PROOF. We may assume that $D^1 = [0, 1]$. Now, if $f : [0, 1] \to [0, 1]$ is a diffeomorphism fixed on $\{0, 1\}$, then $x \mapsto (1 - t) f(x) + tx$ is a diffeomorphism for every $t \in [0, 1]$, because its derivative is always > 0. □

COROLLARY 2.7.B. $\mathrm{Diff}(S^1)$ *is homotopy equivalent to* S^1 *and, moreover, contains* S^1 *as a deformation retract.*

PROOF. Note that S^1 is naturally contained in $\mathrm{Diff}(S^1)$ as the group of rotations. Fix $x \in S^1$ and consider the Serre fibration $\mathrm{Diff}(S^1) \to \mathrm{Emb}(\{x\}, S^1)$. Its base can be obviously identified with S^1, and $S^1 \subset \mathrm{Diff}(S^1)$ defines a section. Using this section, it is easy to see that this Serre fibration is actually a fiber bundle. If we notice that its fiber is homotopy equivalent to $\mathrm{Diff}(D^1 \operatorname{fix} \partial)$, the corollary follows. □

The first basic result about the groups of diffeomorphisms of surfaces is the following theorem of Smale [215].

THEOREM 2.7.C. $\mathrm{Diff}(D^2 \operatorname{fix} \partial)$ *is contractible, where* D^2 *is the* 2*-dimensional disc.*

In that follows, we understand by a *weak deformation retract* of a topological space X a subspace $Y \subset X$ such that the inclusion $Y \to X$ is a weak homotopy equivalence. The following corollary is also due to Smale [215] (who actually proved it with all adjectives "weak" omitted).

COROLLARY 2.7.D. $\mathrm{Diff}(D^2)$ *is weakly homotopy equivalent to a circle.* $\mathrm{Diff}(S^2)$ *is weakly homotopy equivalent to* $\mathrm{SO}(3)$ *and, moreover, contains* $\mathrm{SO}(3)$ *as a weak deformation retract.*

PROOF. Consider the (Serre) fibration $\mathrm{Diff}(D^2) \to \mathrm{Diff}(\partial D^2)$. Its base is homotopy equivalent to S^1 by Corollary 2.7.B and its fiber is homotopy equivalent to a point by Theorem 2.7.C. The first statement follows. Now, fix a disc $D^2 \subset S^2$ and consider the fibration $\mathrm{Diff}(S^2) \to \mathrm{Emb}(D^2, S^2)$. By Theorem 2.6.D its base $\mathrm{Emb}(D^2, S^2)$ is weakly homotopy equivalent to the unit tangent bundle of S^2, which can be identified with $\mathrm{SO}(3)$. The fiber is $\mathrm{Diff}(S^2 \operatorname{fix} D^2)$, which can be identified with $\mathrm{Diff}(S^2 \setminus \operatorname{int} D^2 \operatorname{fix} \partial D^2)$. But $S^2 \setminus \operatorname{int} D^2$ is a disc, hence by Theorem 2.7.C the fiber is contractible. It follows that $\mathrm{Diff}(S^2)$ is weakly homotopy equivalent to $\mathrm{SO}(3)$. Moreover, the fibration $u : \mathrm{Diff}(S^2) \to UT(S^2) = \mathrm{SO}(3)$ from 2.6 is a weak homotopy equivalence. The inclusion of $\mathrm{SO}(3)$ in $\mathrm{Diff}(S^2)$ as the subgroup of rotations defines a section of this fibration. Clearly, the image of this section, i.e., $\mathrm{SO}(3)$, is a weak deformation retract of $\mathrm{Diff}(S^2)$. □

COROLLARY 2.7.E. *Let A be an annulus (i.e., a manifold diffeomorphic to* $S^1 \times [0, 1]$*). Then* $\mathrm{Diff}(A \operatorname{fix} \partial)$ *has weakly contractible components, and the group of components* $\pi_0(\mathrm{Diff}(A \operatorname{fix} \partial))$ *is isomorphic to* $\mathbf{Z}$.

PROOF. Consider the fibration $\mathrm{Diff}(D^2 \operatorname{fix} \partial) \to \mathrm{Emb}(B^2, D^2)$, where D^2 is a disc and B^2 is a disc in the interior of D^2. Its total space is contractible by Theorem 2.7.C. By Theorem 2.6.D its base is weakly homotopy equivalent to the unit tangent bundle of D^2 and, hence, to S^1. Its fiber is $\mathrm{Diff}(D^2 \setminus \operatorname{int} B^2 \operatorname{fix} \partial)$, which can be identified with $\mathrm{Diff}(A \operatorname{fix} \partial)$. The corollary easily follows from these facts and the homotopy sequence of our fibration. □

THEOREM 2.7.F. *Let* $T = S^1 \times S^1$ *be a torus. The components of* $\mathrm{Diff}(T)$ *are weakly homotopy equivalent to T and, in fact, contain T as a weak deformation retract.*

Note that since T is a Lie group, T is naturally contained in $\mathrm{Diff}(T)$. In fact, this subgroup $T \subset \mathrm{Diff}(T)$ is a (weak) deformation retract of the component of the identity of $\mathrm{Diff}(T)$.

THEOREM 2.7.G. *If S is a surface of negative Euler characteristic, then the components of* $\mathrm{Diff}(S\,\mathrm{fix}\,\partial)$ *and of* $\mathrm{Diff}(S)$ *are weakly contractible.*

Theorems 2.7.F and 2.7.G (in fact, with adjectives "weak" omitted) are due to Earle and Eells [46]. The proofs in [46] are analytical and are based on the theory of Beltrami equations. The corresponding results about spaces of homeomorphisms are due to Hamstrom [73,74]; her proofs are purely topological. The corresponding results in the piecewise linear category were obtained by Scott [210]. Nowadays the easiest way to obtain Theorems 2.7.F and 2.7.G is, probably, to consider them as 2-dimensional versions of 3-dimensional results of Hatcher [88] and the author [94] about Haken manifolds (cf. end of 2.5). Unfortunately, such an approach is not written down yet.

THEOREM 2.7.H. *Let S be a surface of negative Euler characteristic and let C be a nontrivial circle on S. Then the components of* $\mathrm{Emb}(C, S)$ *are weakly homotopy equivalent to a circle and the components of* $\mathrm{Sub}(C, S)$ *are weakly homotopy equivalent to a point.*

This theorem can be deduced from the above results about spaces of diffeomorphisms using fibrations from 2.6. Alternatively, if one proves Theorems 2.7.F and 2.7.G along the lines of theorems of [88] and [94] about Haken manifolds, then this theorem is, essentially, the main step of the proof.

Finally, we give an application of these theorems which will be needed later (see 6.3).

THEOREM 2.7.I. *Let R be a connected subsurface of a connected surface S of negative Euler characteristic. Suppose that the boundary of R is connected and nontrivial in S. Then the natural map $\mathcal{M}_R \to \mathcal{M}_S$, given by the extension of diffeomorphisms of R fixed on ∂R by the identity to diffeomorphisms of S, is injective.*

PROOF. Let T be the closure of $S \setminus R$. Clearly, T is connected. The natural map $\mathrm{Diff}(S\,\mathrm{fix}\,\partial) \to \mathrm{Emb}(T, S)$ is a Serre fibration with fiber $\mathrm{Diff}(R\,\mathrm{fix}\,\partial)$. The map $\pi_0(\mathrm{Diff}(R\,\mathrm{fix}\,\partial)) \to \pi_0(\mathrm{Diff}(S\,\mathrm{fix}\,\partial))$ induced by the inclusion of the fiber in the total space is nothing other as our map $\mathcal{M}_R \to \mathcal{M}_S$. In view of the homotopy sequence of this fibration, this map will be injective if $\pi_1(\mathrm{Emb}(T, S)) = 0$. Note that since ∂R is nontrivial, T has negative Euler characteristic and hence $\pi_1(\mathrm{Diff}(T)) = 0$. Let us consider now the natural map $\mathrm{Emb}(T, S) \to \mathrm{Sub}(T, S)$ from Theorem 2.6.B, which is a Serre fibration with fiber $\mathrm{Diff}(T)$. In view of the homotopy sequence of this fibration, $\pi_1(\mathrm{Emb}(T, S)) = 0$ if $\pi_1(\mathrm{Sub}(T, S)) = 0$. But a submanifold of S diffeomorphic to T is determined by its boundary (may be up to two possibilities, if R is diffeomorphic to T). It follows that $\pi_1(\mathrm{Sub}(T, S)) = \pi_1(\mathrm{Sub}(\partial T, S))$. The latter group is 0 in view of Theorem 2.7.H. This completes the proof. □

2.8. *Gluing discs to a surface*

If a surface has nonempty boundary, then we can construct a closed surface or a surface with a smaller number of boundary components by gluing discs to boundary components.

Very often this gluing procedure allows us to reduce theorems about surfaces with possibly nonempty boundary to the case of closed surfaces, or use the induction on the number of boundary components. In this section we discuss some basic constructions and results used in such arguments.

Let R be a surface with nonempty boundary and let Q be the result of gluing discs $D_1, \ldots, D_n$ to n different boundary components of R. Given a diffeomorphism $f: R \to R$ preserving setwise all components of ∂R, there exists an extension $f': Q \to Q$ of f. Such an extension is obviously not unique, but it is unique up to isotopy, because the spaces $\mathrm{Diff}(D_i \operatorname{fix} \partial)$ are connected. Hence, extension of diffeomorphisms from R to Q defines a map $\rho: \mathrm{PMod}_R \to \mathrm{PMod}_Q$. Clearly, ρ is a homomorphism. We are interested in particular in its kernel, and to understand it we need the notion of the pure braid groups of surfaces and the properties of these groups discussed below.

For any surface S, consider $S^{[m]} = \{(y_1, \ldots, y_m) \in S^m\colon\ y_i \neq y_j \text{ for } i \neq j\}$. The *pure braid group* $PB_m(S)$ is defined as $\pi_1(S^{[m]})$, where we tacitly assume that some base point $(x_1, \ldots, x_m) \in S^{[m]}$ have been chosen.

Let $S_{(i)} = S \setminus \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$ (note that $S_{(i)}$ is a noncompact surface). Let x_i be the base point of $S_{(i)}$. The map $S_{(i)} \to S^{[m]}$ given by the formula $x \mapsto (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_m)$ induces a map $\pi_1(S_{(i)}, x_i) \to \pi_1(S^{[m]}) = PB_m(S)$.

THEOREM 2.8.A. *The group $PB_m(S)$ is generated by the images of the above maps $\pi_1(S_{(i)}, x_i) \to \pi_1(S^{[m]}) = PB_m(S)$.*

PROOF. Use induction by m and the natural fibrations $S^{[l]} \to S^{[l-1]}$ given by the formula $(y_1, \ldots y_l) \to (y_1, \ldots, y_{l-1})$. □

We are mainly interested in the group $PB_n(Q)$. We will assume that the base point $(x_1, \ldots, x_n) \in Q^{[n]}$ has been chosen in such a way that $x_i \in \operatorname{int} D_i$, $1 \leqslant i \leqslant n$. Let Q_i be the result of gluing only the disc D_i to R. Clearly, Q_i is contained in $Q_{(i)}$ and hence natural maps $\pi_1(Q_i, x_i) \to PB_n(Q)$ are defined. In fact, Q_i is a deformation retract of $Q_{(i)}$.

COROLLARY 2.8.B. *The group $PB_n(Q)$ is generated by the images of the natural maps $\pi_1(Q_i, x_i) \to PB_n(Q)$.*

Now we construct a homomorphism $j: PB_n(Q) \to \mathrm{PMod}_R$. Let $\beta \in PB_n(Q)$; represent β by a loop in $Q^{[n]}$. Such a loop may be considered as an isotopy of the 0-dimensional submanifold $\{x_1, \ldots, x_n\}$ of Q. Let us extend this isotopy to an isotopy $\{f_t: Q \to Q\}_{0 \leqslant t \leqslant 1}$ of the whole manifold Q; in particular, $f_0 = \mathrm{id}_Q$, $f_1(x_i) = x_i$ for all i. We can easily arrange that, in addition, $f_1(D_i) = D_i$ for all i. (It follows from Theorem 2.6.C that the space of discs in a surface, containing a given point in the interior, is connected.) Then f_1 preserves R and preserves all boundary components of R. We take as $j(\beta)$ the isotopy class of the restriction of f_1 on R. One can easily check that j is correctly defined (using Theorem 2.6.C again). We leave it as an exercise to check that j is a homomorphism. If $n = 1$, then $PB_n(Q) = PB_1(Q)$ is nothing more than the fundamental group $\pi_1(Q)$ and j is a homomorphism $\pi_1(Q) \to \mathrm{PMod}_R$.

The first main property of j is contained in the following theorem.

THEOREM 2.8.C. *The homomorphism ρ is surjective and* $\operatorname{Ker}\rho = \operatorname{Im} j$; *in other words, the sequence*

$$PB_n(Q) \longrightarrow \mathrm{PMod}_R \longrightarrow \mathrm{PMod}_Q \longrightarrow 1$$

is exact. If Q has negative Euler characteristic, then j is injective; in other words, the sequence

$$1 \longrightarrow PB_n(Q) \longrightarrow \mathrm{PMod}_R \longrightarrow \mathrm{PMod}_Q \longrightarrow 1$$

is exact.

PROOF. Let us prove the surjectivity of $\rho : \mathrm{PMod}_R \to \mathrm{PMod}_Q$ first. Represent an element α of PMod_Q by a diffeomorphism $g : Q \to Q$. Connectedness of the spaces of (orientation-preserving) embeddings of discs into a surface easily implies that g is isotopic to a diffeomorphism g' such that $g'(D_i) = D_i$, $1 \leqslant i \leqslant n$. Let γ be the isotopy class of the restriction of g' to R. Clearly, $\rho(\gamma) = \alpha$. This proves the surjectivity of ρ.

Now, let $\gamma \in \operatorname{Ker}\rho$. Let us represent γ by a diffeomorphism $f : R \to R$ and extend f to a diffeomorphism $f' : Q \to Q$. We may assume that $f'(x_i) = x_i$, $1 \leqslant i \leqslant n$. Since $\gamma \in \operatorname{Ker}\rho$, the diffeomorphism f' is isotopic to the identity id_Q. Let $\{f'_t : Q \to Q\}_{0\leqslant t\leqslant 1}$ be some isotopy between $f' = f'_0$ and $f'_1 = \mathrm{id}_Q$. Then $t \mapsto (f'_t(x_1), \ldots, f'_t(x_n))$ defines a loop in $Q^{[n]}$ with the base point $(x_1, \ldots, x_n)$ and, hence, an element $\beta \in \pi_1(Q_{[n]}) = PB_n(Q)$. Obviously, $j(\beta) = \gamma$. This proves that $\operatorname{Ker}\rho \subset \operatorname{Im} j$. The opposite inclusion follow directly from the definitions.

Suppose now that the Euler characteristic of Q is negative. Let $\beta \in \operatorname{Ker} j$. Let us represent β by a loop in $Q^{[n]}$ and construct an isotopy $\{f_t : Q \to Q\}_{0\leqslant t\leqslant 1}$ as in the definition of j. In particular, $f_1(D_i) = (D_i)$ for all i. Since $\beta \in \operatorname{Ker} j$, the restriction of f_1 on R is isotopic to id_R. It follows that after changing the isotopy, if necessary, we may assume that the restriction of f_1 on R is equal to id_R. Next, in view of Theorem 2.7.C, we may assume that $f_1 = \mathrm{id}_Q$. Now, $\{f_t : Q \to Q\}_{0\leqslant t\leqslant 1}$ defines a loop in $\mathrm{Diff}(Q)$. By Theorem 2.7.G this loop is contractible. Clearly, this implies that the original loop representing β is contractible, and hence $\beta = 1$. □

This theorem is due to Birman [14]. See also [16, Theorem 4.2]. In general, $\operatorname{Ker} j$ is isomorphic to the quotient of the group $PB_n(Q)$ by its center, which can be easily identified; see [16, Chapter 4].

Now, let $\alpha \in \mathrm{Mod}_R$. Suppose that some (and then every) diffeomorphism $f : R \to R$ representing α preserves (setwise) the union of circles ∂D_i, $1 \leqslant i \leqslant n$. Then f can be extended to a diffeomorphism $f' : Q \to Q$. Such extension is unique up to isotopy (cf. the construction of ρ). We may assume that $f'(x_i) = x_{\sigma(i)}$ for some permutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. Extension f' is unique up to isotopy even in the class of such diffeomorphisms (this follows from the fact that discs are connected and the results of 2.6). Such an extension f' induces a map $Q^{[n]} \to Q^{[n]}$ mapping the base point $* = \{x_1, \ldots, x_n\}$ to $\sigma(*) = \{x_{\sigma(1)}, \ldots, x_{\sigma(n)}\}$. In particular, f' induces a homomorphism of the fundamental groups $\pi_1(Q_{[n]}, *) \to \pi_1(Q_{[n]}, \sigma(*))$. It is easy to see that this homomorphism depends only on α. By a slight abuse of notations, we will denote it, even if $\sigma \neq \mathrm{id}$, by

$\alpha_* : PB_n(Q) \to PB_n(Q)$ (keeping in mind that the second pure braid group is defined using a different base point).

Note that, obviously, α_* maps the image of $\pi_1(Q_i, x_i)$ in $PB_n(Q)$ to the image of $\pi_1(Q_{\sigma(i)}, x_{\sigma(i)})$.

The homomorphisms α_* and j are related by the following fundamental formula:

$$\alpha j(\beta)\alpha^{-1} = j(\alpha_*(\beta)),$$

which follows immediately from the constructions of α_* and j.

As the last application of the extension of diffeomorphisms from R to Q, consider the situation when only one disc, D_1, is added to R. Then the extension of diffeomorphisms defines a map $\mathrm{Mod}'_R \to \pi_0(\mathrm{Diff}(Q \operatorname{fix} x_1))$, where Mod'_R is the subgroup of Mod_R consisting of isotopy classes of diffeomorphisms preserving ∂D_1. It turns out that this map is an isomorphism. (Again, this follows from 2.6.)

2.9. *The Dehn–Nielsen–Baer theorems*

For a group π, let $\mathrm{Aut}(\pi)$ be the group of automorphisms of π, and let $\mathrm{Inn}(\pi)$ be the subgroup of inner automorphisms of π. A trivial check shows that $\mathrm{Inn}(\pi)$ is a normal subgroup of $\mathrm{Aut}(\pi)$. The quotient group $\mathrm{Aut}(\pi)/\mathrm{Inn}(\pi)$ is denoted by $\mathrm{Out}(\pi)$ and called the *outer automorphisms group* of π. These groups naturally show up when one deals with homotopy self-equivalences of a (connected) space X with $\pi_1(X) = \pi$. Namely, any self-map $f : X \to X$ defines a homomorphism $f_* : \pi_1(X) \to \pi_1(X)$ and if f is a homotopy equivalence, then f_* is an isomorphism. But, if we do not assume a base point to be fixed, f_* is well defined only up to an inner automorphism of $\pi_1(X) = \pi$. Hence, a homotopy equivalence $f : X \to X$ defines an element $f_* \in \mathrm{Out}(\pi)$. Obviously, f_* depends only on the homotopy class of f.

Now, let $X = S$ be a compact (connected) orientable surface. In view of the above remarks, there is a natural homomorphism $\mathrm{Mod}^{\diamond}_S \to \mathrm{Out}(\pi_1(S))$. If S has nonempty boundary, then the elements of the image of this homomorphism obviously preserve the peripheral structure of $\pi_1(S)$ from 2.2.

THEOREM 2.9.A. *If S is closed and is not a sphere, then the natural homomorphism* $\mathrm{Mod}^{\diamond}_S \to \mathrm{Out}(\pi_1(S))$ *is an isomorphism. If S has nonempty boundary and negative Euler characteristic (i.e., S is not a disc or an annulus), then the natural homomorphism* $\mathrm{Mod}^{\diamond}_S \to \mathrm{Out}(\pi_1(S))$ *is an isomorphism onto the subgroup of* $\mathrm{Out}(\pi_1(S))$ *consisting of elements preserving the peripheral structure.*

One can complement this theorem by a description of the subgroup of $\mathrm{Out}(\pi_1(S))$ corresponding to Mod_S. Suppose first that S is closed (and is not a sphere). (For a review of the cohomology of groups, used below, see 6.1.) Then S is an Eilenberg–MacLane space, and hence $H^2(\pi_1(S), \mathbf{Z}) = H^2(S, \mathbf{Z}) = \mathbf{Z}$. The group $\mathrm{Out}(\pi_1(S))$ naturally acts on $H^*(\pi_1(S), \mathbf{Z})$, because inner automorphisms act trivially on cohomology groups. Let $\mathrm{Out}^+(\pi_1(S))$ be the subgroup of $\mathrm{Out}(\pi_1(S))$ consisting of elements acting trivially on $H^2(\pi_1(S), \mathbf{Z})$. Clearly, the image of Mod_S under the homomorphism $\mathrm{Mod}^{\diamond}_S \to \mathrm{Out}(\pi_1(S))$

is contained in $\mathrm{Out}^+(\pi_1(S))$. If S has nonempty boundary (and negative Euler characteristic), then the image of Mod_S in $\mathrm{Out}(\pi_1(S))$ is obviously contained in the subgroup of $\mathrm{Out}(\pi_1(S))$ consisting of elements preserving the *oriented* peripheral structure from 2.2.

THEOREM 2.9.B. *If S is closed and is not a sphere, then the natural homomorphism* $\mathrm{Mod}_S \to \mathrm{Out}^+(\pi_1(S))$ *is an isomorphism. If S has nonempty boundary and negative Euler characteristic, then the natural homomorphism* $\mathrm{Mod}_S \to \mathrm{Out}(\pi_1(S))$ *is an isomorphism onto the subgroup of* $\mathrm{Out}(\pi_1(S))$ *consisting of elements preserving the oriented peripheral structure.*

For closed S, the surjectivity part of these theorems is due to Dehn, who did not publish his proof, and to Nielsen [191], who published a proof partially based on Dehn's ideas. Another proof was suggested by Seifert [211]. The injectivity part is due to Baer [3]. The surjectivity part for surfaces with boundary was proved by Magnus [148], and the injectivity part – much later – by Zieschang [244]. Proofs and a detailed discussion of these results can be found in [245]; see [245], Theorem 3.3.11 (surjectivity for closed surfaces following Seifert), Theorem 5.7.1 (surjectivity for surfaces with boundary) and Theorem 5.13.1 (injectivity).

In principle, these results give a purely algebraic description of mapping class groups and allow one to reduce every question about them to a purely algebraic question. Surprisingly, this reduction turned out to be not very successful. Probably, the only algebraic property of the mapping class groups which for a long time could be proved only in terms of the outer automorphisms groups is the residually finiteness.

THEOREM 2.9.C. *The groups* Mod°_S *are residually finite, i.e., for every nontrivial element* $f \in \mathrm{Mod}^{\circ}_S$, $f \neq 1$, *there is a homomorphism* $h : \mathrm{Mod}^{\circ}_S \to G$ *onto a finite group G such that* $h(f) \neq 1$.

This theorem is due to Grossman [68], whose proof was based on the description of mapping class groups in terms of outer automorphisms groups and fairly complicated combinatorial group theory arguments. A more conceptual proof, still based on the description in terms of outer automorphisms groups was given by Bass and Lubotzky [6]. Now a simple direct proof, not based on the results of Dehn–Nielsen–Baer, is available; see [107] or [108, Exercise 1].

In one case, namely, if S is a closed torus, the Dehn–Nielsen–Baer theorems provide a complete and efficient description of Mod_S and Mod°_S. In this case $\pi_1(S)$ is Abelian and, moreover, is isomorphic to $\mathbf{Z}^2$, and hence $\mathrm{Out}(\pi_1(S)) = \mathrm{Aut}(\pi_1(S)) = \mathrm{Aut}(\mathbf{Z}^2) = \mathrm{GL}_2(\mathbf{Z})$. This chain of isomorphisms clearly maps the subgroup of $\mathrm{Out}^+(\pi_1(S))$ to $\mathrm{SL}_2(\mathbf{Z})$. It follows that $\mathrm{Mod}^{\circ}_S = \mathrm{GL}_2(\mathbf{Z})$ and $\mathrm{Mod}_S = \mathrm{SL}_2(\mathbf{Z})$ for a closed torus S.

3. Complexes of curves

Complexes of curves are, probably, the most fundamental geometric objects on which the mapping class groups act (but the Teichmüller spaces are, certainly, the most important). They were discovered by Harvey [86,87], and have played an ever increasing role since then. In this section we give the definition and discuss the basic properties of them. We prove that complexes of curves are connected, and do this in a way which can be generalized to prove much stronger homotopy properties of them. After this, we discuss the homotopy type and the hyperbolic properties of them. Later sections will amply illustrate the usefulness of complexes of curves.

3.1. *Definitions*

Complexes of curves are simplicial complexes in the sense of [216, Chapter 3] or [28], for example. Thus, a *simplicial complex* consists of a set of *vertices* and a set of *simplices*. Simplices are nonempty finite sets of vertices, subject only to the following two conditions: a nonempty subset of a simplex is a simplex; every vertex belongs to some simplex. The *dimension* of a simplex is the number of vertices in it minus 1. One-dimensional simplices play a special role and are called *edges*.

The vertices of the *complex of curves* $C(S)$ of a compact orientable surface S are the isotopy classes of simple closed curves (as usual, we call them *circles*) on S, which are *non-trivial* in the sense of 2.3, i.e., not contractible in S into a point or into the boundary ∂S. We denote the isotopy class of a circle C by $\langle C\rangle$. A set of vertices $\{\gamma_0, \dots, \gamma_n\}$ is declared to be a simplex if and only if $\gamma_0 = \langle C_0\rangle, \dots, \gamma_n = \langle C_n\rangle$ for some pairwise disjoint circles $C_0, \dots, C_n$. The extended mapping class group $\mathrm{Mod}_S^\diamond$ acts on $C(S)$ in an obvious way: the isotopy class $f \in \mathrm{Mod}_S$ of a diffeomorphism $F : S \to S$ maps $\langle C\rangle$ to $\langle F(C)\rangle$.

$C(S)$ can be defined in another way. According to 2.5, we can equip S with a Riemannian metric of constant curvature with geodesic boundary. Then we can take as the set of vertices of $C(S)$ the set of geodesic circles in $S \setminus \partial S$ and as simplices the sets of pairwise disjoint geodesic circles. In view of Lemma 2.5.A, this definition is equivalent to the previous one. This definition makes clear that a set of vertices is a simplex if and only if all its 2-element subsets are simplices (i.e., any two its vertices are connected by an edge). In the language of the theory of buildings (cf. [28], for example) this means that $C(S)$ is a *flag complex*. Thus, $C(S)$ is, in fact, completely determined by its 1-skeleton. Higher-dimensional simplices do not contribute any new combinatorial information, but they make the homotopy theory of $C(S)$ much more relevant and interesting.

Speaking about the homotopy properties of $C(S)$ we, of course, have in mind the homotopy properties of its *geometric realization* (cf. [216]), which is a topological space and even a CW-complex.

Obviously, $C(S)$ is an infinite complex. It is even *locally infinite* when it has positive dimension: every vertex is connected by an edge with an infinite number of vertices. But it is finitely-dimensional: its dimension is equal to the maximal number of pairwise non-isotopic disjoint non-trivial circles on S minus 1. It follows that if S is a connected orientable surface of genus g with b boundary components, then the dimension of $C(S)$ is equal to $3g - 4 + b$, except in the case $g = 1$ and $b = 0$, when the dimension is zero (if this formula leads to a negative number, the complex of curves is empty).

3.2. *Connectivity of $C(S)$ and the Morse–Cerf theory*

Before discussing other homotopy properties of $C(S)$, we will present a proof of the connectedness of $C(S)$. This proof can be generalized to give the best possible connectivity result for closed surfaces S and surfaces S with one boundary component; cf. Theorem 3.2.C (the case of surfaces with more boundary components can be reduced to the case of closed surfaces in at least two different ways; cf. 3.3). It is based on some elementary facts from the theory of singularities, more precisely, on a generalization of the Morse theory to families of functions due to Cerf [32,33]. Namely, we need the following result.

LEMMA 3.2.A. *Any family of functions $\{f_t : S \to \mathbf{R}\}_{0 \leqslant t \leqslant 1}$ can be approximated (say, in the C^∞-topology) by a family of functions $\{g_t\}_{0 \leqslant t \leqslant 1}$ such that all functions g_t belong to one of the following three classes:*

(i) *Morse functions with all* critical values (*i.e., values at critical points*) *different*;

(ii) *Morse functions with two equal critical values and all other critical values different from them and from each other*;

(iii) *functions having all critical values different and exactly one non-Morse critical point such that in appropriate local coordinates (x, y) around this critical point the function has the form $x^3 \pm y^2 + c$ for some constant $c \in \mathbf{R}$.*

This lemma was originally proved in [32, Chapter II], and has become well known since then. The proof shows that one can also require that all but finitely many functions g_t belong to the first class, but we do not need this fact. For a general introduction to the relevant ideas of the theory of singularities one can recommend a nice short book by Poston and Stewart [204]; cf. [204, Chapters 1, 2].

THEOREM 3.2.B. *Suppose that S is neither a sphere with $\leqslant 4$ holes, nor a torus with $\leqslant 1$ holes. Then $C(S)$ is connected.*

PROOF. If S is a torus with two holes, it is easy to see directly that $C(S)$ is connected. We leave this as an exercise to the reader, with the following hints: there are two types of non-trivial circles, namely, nonseparating circles and circles separating S into a torus with one hole and a disc with two holes; the circles of the first type are never disjoint, but two such circles intersecting at exactly one point can be made (by an isotopy) disjoint from a circle of the second type.

In the remaining part of the proof we assume that S is neither a sphere with $\leqslant 4$ holes, nor a torus with $\leqslant 2$ holes.

Let $\gamma = \langle D \rangle$, $\gamma' = \langle D' \rangle$ be two vertices of $C(S)$. We need to prove that there is a sequence of vertices $\gamma = \gamma_0, \dots, \gamma_n = \gamma'$ such that γ_i is connected by an edge with γ_{i+1} (or is equal to γ_{i+1}) for all $i = 0, \dots, n-1$.

To begin with, let us choose two smooth functions $f, f' : S \to \mathbf{R}$ such that D (respectively D') is a component of a level set $f^{-1}(a)$ of f (respectively, of f') and, moreover, f and f' have no critical points on D and D', respectively. In addition, we may assume that both f and f' are Morse functions having different critical values. Clearly, we can connect f and f' by a path $\{f_t : S \to \mathbf{R}\}$, $0 \leqslant t \leqslant 1$, in the space of smooth functions, $f = f_0$, $f' = f_1$.

Let $\{g_t\}_{0 \leqslant t \leqslant 1}$ be some approximation of $\{f_t\}_{0 \leqslant t \leqslant 1}$ provided by Lemma 3.2.A. Since g_0 is C^∞-close to f_0, some component C of the level set $g^{-1}(a)$ is a circle close and isotopic to D. In particular, this circle belongs to the isotopy class γ. Similarly, some component

C' of some level set of g_1 belongs to the isotopy class γ'. The following claim provides similar components of level sets for all functions g_t.

CLAIM. *If a function $g: S \to \mathbf{R}$ belongs to one of the three classes of Lemma* 3.2.A, *then some component of some level set $g^{-1}(r)$ of g contains no critical points and is a* non-trivial *circle.*

In order to prove this claim, let us choose a critical point x of g which is not a local maximum or minimum (such a critical value exists, because S is not a sphere or a disc). Let $c = g(x)$ and let L be the component of the level set $g^{-1}(c)$ containing x. Choose an $\varepsilon > 0$ such that c is the only critical value in $[c - \varepsilon, c + \varepsilon]$. Then exactly one component of the set $g^{-1}([c-\varepsilon, c+\varepsilon])$ contains L. Let us denote this component by L_ε. If one of the boundary components of L_ε is non-trivial, then we are done. Suppose that all of them are trivial, i.e., bound in S either a disc, or an annulus together with a boundary component of S.

If one of these discs or annuli contains L_ε, then we replace g by a function g' equal to g outside this disc (or annulus) X and having only one Morse critical point on it if it is a disc (or no critical points if it is an annulus). We may assume that if g' has a critical point in X (i.e., if X is a disc), then the corresponding critical value is different from all other critical values of g'. Note that if X is a disc, then g has at least two critical points on X (the point x and a minimum or a maximum), and g' has only one. If X is an annulus, then g has at least one critical point on X (namely, the point x), and g' has none. In both cases g' has fewer critical points than g. If some component of some level set of g' is non-trivial, then the same is true for g itself, because all components of level sets of g' contained in X are trivial. So, in this case our problem is reduced to a similar problem for the function g', which has fewer critical points than g. Since g' obviously belongs to one of three classes of Lemma 3.2.A, we can use induction.

If none of these discs or annuli contain L_ε, then S is equal to the union of L_ε with these discs or annuli. It follows that the Euler characteristic $\chi(S)$ is equal to $\chi(L_\varepsilon) + d$, where $d \geqslant 0$ is the number of the discs. Clearly, L is a deformation retract of L_ε and, hence, $\chi(S) = \chi(L) + d$. But, $\chi(L) = -1$ if L contains one Morse critical point, $\chi(L) = -2$ if L contains two Morse critical points and $\chi(L) = 0$ if L contains a non-Morse critical point (in the last case L does not contain any other critical points and is a topological circle with a cusp at the critical point). The various possibilities for L are presented in Figure 2. It follows that $\chi(S) \geqslant -2$ in this case. But any such S is either excluded by the assumptions of the theorem, or is a torus with two holes, which is also excluded now. This completes the proof of the claim.

Now we return to the proof of the theorem. The above claim implies that for any t, $0 \leqslant t \leqslant 1$, the function g_t has a component C_t of a level set $g_t^{-1}(a_t)$ such that C_t does not contain any critical point of g_t and is a non-trivial circle. We may assume that $C_0 = C$ and $C_1 = C'$. If u is sufficiently close to t, say u belongs to a neighborhood U_t of t, then the level set $g_u^{-1}(a_t)$ of the function g_u has a component C_{tu} close and isotopic to C_t, because g_t has no critical points on C_t. The family of neighborhoods $\{U_t\}_{0 \leqslant t \leqslant 1}$ forms an open cover of the interval $[0, 1]$. By using a well known Lebesgue lemma we can divide the interval $[0, 1]$ by several points $0 = x_0 \leqslant x_1 \leqslant \cdots \leqslant x_n = 1$ into intervals $[x_i, x_{i+1}]$

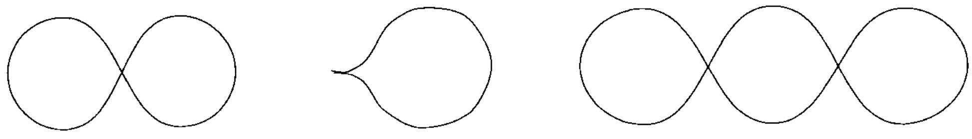


Fig. 2.

such that any interval $[x_i, x_{i+1}]$, $0 \leqslant i \leqslant n-1$, is contained in a neighborhood U_{t_i}. We may assume that $t_0 = 0$ and $t_{n-1} = 1$.

Let $\gamma_i = \langle C_{t_i} \rangle$. We claim that $\gamma_0, \dots, \gamma_n$ is the required sequence of vertices. By our choice, $\gamma_0 = \gamma$ and $\gamma_n = \gamma'$. Let us prove that γ_i is connected by an edge with γ_{i+1} for $i = 0, \dots, n-1$. Let $0 \leqslant i \leqslant n-1$, and let $x = x_i$, $y = x_{i+1}$, $z = x_{i+2}$ and $v = t_i$, $w = t_{i+1}$. Note that C_v is isotopic to C_{vy} because $y \in [x, y] \subset U_v$ and C_w is isotopic to C_{wy} because $y \in [y, z] \subset U_w$. Hence, $\gamma_i = \langle C_{t_i} \rangle = \langle C_v \rangle = \langle C_{vy} \rangle$ and $\gamma_{i+1} = \langle C_{t_{i+1}} \rangle = \langle C_w \rangle = \langle C_{wy} \rangle$. But, C_{vy} and C_{wy} are components of the level sets of the *same* function g_y. Hence, they are either equal or disjoint and their isotopy classes γ_i and γ_{i+1} are either equal or connected by an edge. This completes the proof. □

This theorem for closed surfaces S was originally deduced by the author [96] from the results of Hatcher and Thurston [91], which were also based on the ideas of the Morse–Cerf theory. Later on, the method was streamlined and extended to cover the higher connectivity of $C(S)$; cf. the discussion below. In the meantime, Harer [77,78] proved that $C(S)$ is homotopy equivalent to a bouquet of spheres, all having the same dimension, and computed this dimension, giving, in particular, the best possible estimate of the higher-connectedness. Cf. 3.3.

In [32] Cerf also proved a version of Lemma 3.2.A for 2-parameter families of functions. Here, some new types of functions appear, such as Morse functions with three equal critical values or functions having critical points of the form $\pm x^4 \pm y^2 + c$. Using this version of Lemma 3.2.A, one can modify the proof of Theorem 3.2.B and prove the simply connectivity of $C(S)$ in most cases. Instead of a subdivision of the interval $[0, 1]$ into subintervals, one should use a sufficiently fine triangulation of a disc.

The Morse–Cerf theory is based on a complete classification of possible singularities of functions in *generic* families. Such a classification is unknown for families with sufficiently many parameters. But, in fact, one can carry out the main arguments of the proof of Theorem 3.2.B without complete classification. One needs to prove only some restrictions on the possible complexity of singularities of functions in generic families (cf. [98, Lemmas 2.3 and 2.4]). This program is carried out in [98], Sections 1 and 2 and leads to the following result.

THEOREM 3.2.C. *If S is a closed surface (respectively, a surface* 1 *boundary component, a surface with $\geqslant 2$ boundary components), then the geometric realization of $C(S)$ is $-\chi(S) - 1 = 2g - 3$-connected (respectively, $-\chi(S) - 2$-connected, $-\chi(S) - 3$-connected), where $\chi(S)$ is the Euler characteristic of S.*

This result is not the best possible for surfaces with $\geqslant 2$ boundary components (the best is the $-\chi(S)-2$-connectedness), but the closed case is sufficient for the main applications (cf., for example, 6.4), and, in fact, the general case can be reduced to it.

3.3. *The homotopy type of the complexes of curves*

The following theorem gives an almost complete description of the homotopy type of $C(S)$.

THEOREM 3.3.A. *The geometric realization of $C(S)$ is homotopy equivalent to a bouquet of spheres of dimension $-\chi(S)$ if S is closed and of dimension $-\chi(S)-1$ if S has nonempty boundary, where $\chi(S)$ is the Euler characteristic of S.*

This theorem is due to Harer [78]. His proof is based on a connection of the complexes of curves with ideal triangulations of Teichmüller spaces (cf. 5.5) and a combinatorial argument used to give an upper estimate of the *homotopy dimension* of $C(S)$, i.e., of the minimal dimension of a CW-complex homotopy equivalent to it. This upper estimate, in fact, matches the connectivity results, and together they imply that $C(S)$ is homotopy equivalent to a bouquet of spheres. Another proof, based on Theorem 3.2.C, a duality between the properties of $C(S)$ and Mod_S (cf. 6.1 and 5.4) and Harer's combinatorial argument giving an upper estimate of homotopy dimension was suggested by the author [98]. We refer to [98] for a detailed presentation of this proof and a discussion of its relations with Harer's approach; cf. [98, Section 6].

Both approaches use an induction on the number of boundary components reducing, in fact, the general case to the closed one. Harer's inductive argument [78] is made on the geometric level of complexes of curves. He shows that making a hole in a surface with nonempty boundary increases the connectivity of $C(S)$ by 1 (making the first hole is more complicated). In [98], this inductive argument is made on the algebraic level of the mapping class groups, where the exact sequences of 2.8 together with theory of groups with duality [13] provide a convenient tool. An improved version of this argument is used in the proof of Theorem 6.4.B.

Table 1, summarizing the dependence of the homotopy dimension $\mathrm{hdim}\, C(S)$ and the usual dimension $\dim C(S)$ of $C(S)$ on the genus g and the number b of boundary components of S, is often quite useful.

Table 1

g	b	$\mathrm{hdim}\, C(S)$	$\dim C(S)$
$g=0$	$b\leqslant 3$	-1	-1
$g=0$	$b\geqslant 3$	$b-4$	$b-4$
$g=1$	$b=0$	0	0
$g=1$	$b\geqslant 1$	$b-1$	$b-1$
$g\geqslant 2$	$b=0$	$2g-2$	$3g-4$
$g\geqslant 2$	$b\geqslant 1$	$2g-3+b$	$3g-4+b$

COROLLARY 3.3.B. *Suppose that S is connected and has nonempty boundary. Then the structure of the simplicial complex $C(S)$ (considered up to isomorphism) determines if the*

genus of S is $\leqslant 1$*, and the whole topological type of S if it turns out that the genus of S is* $\geqslant 2$.

PROOF. The table shows that $\operatorname{hdim} C(S) = \dim C(S)$ if and only if the genus is $\leqslant 1$ or the genus is 2 and the boundary is empty. The last line of the table allows us to find g and b if we know already that $g \geqslant 2$ and $b \geqslant 1$. □

3.4. *Hyperbolic properties*

Now we turn from the homotopy properties to the geometry of complexes of curves. Let $C_1(S)$ be the 1-skeleton of $C(S)$, i.e., the graph with the same vertices and edges as $C(S)$. As we noted in 3.1, $C(S)$ is a flag complex and, hence, $C_1(S)$ completely determines the whole structure of $C(S)$.

Any graph or, rather, its geometric realization, has a canonical structure of a metric space, which is constructed as follows: any edge is made isometric to the interval [0, 1] and then the distance between two points is defined to be the infimum of the lengths of paths connecting them. Clearly, the distance between two vertices γ, γ' is equal to the length n of the shortest chain $\gamma = \gamma_0, \dots, \gamma_n = \gamma'$ such that γ_i is connected by an edge with γ_{i+1} (compare the proof of Theorem 3.2.B). This structure of a metric space turns any graph into a *geodesic space*. This means that any two points x, y can be connected by an isometric image of an interval in the real line $\mathbf{R}$. Any such image is called a *geodesic* and is denoted by $[xy]$, despite the fact that it is usually not unique.

A geodesic metric space X is called *δ-hyperbolic*, where $\delta \geqslant 0$ is some real number, if for any three points $x, y, z \in X$ any geodesic $[xz]$ is contained in a δ-neighborhood of the union $[xy] \cup [yz]$. A geodesic metric space is called *hyperbolic* if it is δ-hyperbolic for some $\delta \geqslant 0$. These notions are due to Gromov [67] and serve as the starting point of his theory of hyperbolic spaces and groups. The main examples are provided by the classical hyperbolic spaces, (infinite) trees (i.e., graphs without cycles), and Cayley graphs of fundamental groups of closed negatively curved manifolds. (For trees one can take $\delta = 0$.) Note also that any geodesic metric space of finite diameter D is trivially δ-hyperbolic with $\delta = D$.

THEOREM 3.4.A. *If $C_1(S)$ is connected, then its geometric realization is a hyperbolic space of infinite diameter.*

This remarkable and completely unexpected theorem was recently proved by Masur and Minsky [156]. An overview of the proof is provided by Minsky [165]. The replacement of $C(S)$ by $C_1(S)$ is not significant: we can turn the geometric realization of $C(S)$ into a metric space by making every simplex be isometric to a regular simplex with edges of length 1. Then the resulting metric space will also be hyperbolic. This easily follows from Theorem 3.4.A and basic properties of hyperbolic spaces. It simply seems that it is more convenient to work with $C_1(S)$ in this context.

Even the fact that $C_1(S)$ is of infinite diameter is non-trivial and interesting. It means that for any natural number N there are two vertices of $C(S)$ which cannot be connected by a chain of edges shorter than N. This seems to be nearly obvious, but, apparently, was not proved before [156].

The fact that $C(S)$ is not locally finite presents a serious obstacle if one attempts to use Theorem 3.4.A along the lines of Gromov's theory of hyperbolic groups and spaces [67]. In [157], Masur and Minsky developed some tools to overcome this obstacle. As a first application to the mapping class groups, they proved the following theorem about pseudo-Anosov elements (for the definition, see 7.1) of the mapping class groups.

THEOREM 3.4.B. *Let us fix a finite set of generators of* Mod_S *and let us denote by* $|\cdot|_W$ *the minimal word length with respect to these generators. If* f_1, f_2 *are two conjugate pseudo-Anosov elements of* Mod_S, *then* $f_1 = g f_2 g^{-1}$ *for an element* $g \in \mathrm{Mod}_S$ *such that*

$$|g|_W \leqslant C\big(|f_1|_W + |f_2|_W\big),$$

where the constant C depends only on S and the generating set.

To put this theorem into proper perspective, one should note that Hemion [92] proved that the conjugacy problem for the mapping class group is (algorithmically) solvable. Pseudo-Anosov elements are, in a definite sense, typical elements of Mod_S, and Mosher [180] gave an explicit algorithm for determining the conjugacy for pseudo-Anosov elements. In both [92] and [180] no explicit bounds on the time required by the algorithm was given. Theorem 3.4.B obviously implies an explicit bound for the conjugacy problem, because it provides a bound on the word length of a conjugating element.

4. Dehn twists, generators, and relations

The main examples of nontrivial elements of mapping class groups, and also the main building blocks in various constructions of other elements, are the so-called Dehn twists. They were introduced for the first time by Dehn [44], and then played a central role in his fundamental paper [45]. This section starts with a definition and some basic properties of Dehn twists, and then proceeds to one of their main applications, namely, to their role as generators of mapping class groups. The material of this section is crucial for developing an intuitive feeling of the mapping class groups.

In this section S is always assumed to be a compact *oriented* surface.

4.1. *Dehn twists*

The goal of this section is to define Dehn twists and to prove their basic properties. Dehn twists are, by definition, the isotopy classes of some special diffeomorphisms called twist diffeomorphisms. We begin with a discussion of the latter.

We start with a description of a standard twist diffeomorphism of an annulus. Let A be the annulus in $\mathbf{R}^2$ given by the inequality $1 \leqslant r \leqslant 2$ in the standard polar coordinates (r, θ) in $\mathbf{R}^2$. Its boundary ∂A consists of two components $\partial_1 A$, $\partial_2 A$ defined by the equations $r = 1$, $r = 2$ respectively. Let us fix a smooth function $\varphi : \mathbf{R} \to \mathbf{R}$ such that $\varphi(x) = 0$ for

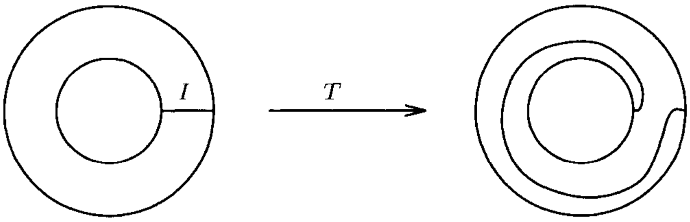


Fig. 3.

$x \leqslant 1$, $\varphi(x) = 2\pi$ for $x \geqslant 2$, $0 \leqslant \varphi \leqslant 2\pi$, and the derivative φ' is $\geqslant 0$. Let us define a diffeomorphism $T : A \to A$ by the formula $T(r, \theta) = (r, \theta + \varphi(r))$ in our polar coordinates (r, θ). We call T the *standard twist diffeomorphism* of the annulus A. It is easy to see that, up to an isotopy fixed on the boundary ∂A, the diffeomorphism T does not depend on the choice of φ. Figure 3 illustrates the action of T. In fact, up to an isotopy fixed on the boundary, T is the unique diffeomorphism fixed on the boundary and taking the radial arc I in Figure 3(a) into an arc isotopic to the arc in Figure 3(b). (Using the fact that $\mathrm{Diff}(I \operatorname{fix} \partial I)$ is connected, any two such diffeomorphisms can be made equal on I by an isotopy. By an additional small isotopy, these two diffeomorphisms can be made equal on a neighborhood N of $I \cup \partial A$. Consider a disc D contained in A with the boundary contained in N. Since $\mathrm{Diff}(D \operatorname{fix} \partial)$ is connected by Theorem 2.7.C, any two such diffeomorphisms can be made equal by an isotopy.) So, Figure 3 defines T up to an isotopy fixed on ∂A.

Now, let $e : A \to S$ be an *orientation-preserving* embedding of A into our surface S. We may use e to transplant T from A to S as follows: take the diffeomorphism $e \circ T \circ e^{-1} : e(A) \to e(A)$ and extend it by the identity to a diffeomorphism $T_e : S \to S$. We call any such diffeomorphism T_e a *twist diffeomorphism* of S. Clearly, up to isotopy fixed on ∂S, the diffeomorphism T_e depends only on the isotopy class of the embedding e. The isotopy class of an embedding $e : A \to S$ is determined by the isotopy class of the (oriented) image $e(a)$ of the *oriented* (say, counterclockwise) axis $a = \{(r, \theta) \colon r = 3/2\}$ of the annulus A and by which boundary component of A is mapped to the right side of $e(a)$ (the notion of the *right side* is determined by the orientations of $e(a)$ and S). So, given the isotopy class of the *unoriented* image $e(a)$, there are four possible isotopy classes of e, but only two of them are orientation-preserving. In fact, if e is orientation-preserving, then e has to map $\partial_2 A$ to the right of $e(a)$. It is easy to see (draw a picture!) that both orientation-preserving embeddings e with the same isotopy class of $e(a)$ lead to isotopic diffeomorphisms T_e.

In particular, up to isotopy fixed on ∂S, the diffeomorphism T_e depends only on the image $C = e(a)$ of the axis a of A, and even only on its isotopy class $\gamma = \langle C \rangle$. By a slight abuse of the language, we call this diffeomorphism a twist diffeomorphism *about* C. Clearly, any twist diffeomorphism preserves orientation and is fixed on the boundary of S. We call the isotopy class of a twist diffeomorphism about a circle C the *Dehn twist about* C. Since this isotopy class depends only on the isotopy class $\gamma = \langle C \rangle$ of C, we usually denote it by t_γ and call it also the *Dehn twist about* γ. There is an ambiguity in these notations and terminology; namely, one may consider only isotopies fixed on the boundary and then $t_\gamma \in \mathcal{M}_S$, or one may consider all isotopies and then $t_\gamma \in \mathrm{PMod}_S$. The context usually resolves this ambiguity.

The Dehn twists we defined are often called the *left* Dehn twists, and the elements of the form t_γ^{-1} are called the *right* Dehn twists. Similarly, one may speak about the *left* and the *right* twist diffeomorphisms. This terminology is justified, first, by the fact that the right twist diffeomorphisms and Dehn twists can be constructed in a manner completely similar to the construction of the left ones, but using orientation-reversing embeddings $A \to S$ instead of orientation-preserving ones (we leave this as an exercise). Second, one can distinguish between the right and the left twist diffeomorphisms in the following way. Let $e: A \to S$ be an embedding and T_e be the corresponding (left or right) twist diffeomorphism. If J is an arc in S transversely intersecting $C_1 = e(\partial_1 A)$, then as we approach the annulus $e(A)$ along $T_e(J)$ and pass the circle C_1, we will be turning left if T_e is a left twist diffeomorphism and we will be turning right if T_e is a right twist diffeomorphism. Of course, the meaning of turning left or right is derived from the orientation of S. A useful exercise is to convince yourself that we will turn to the same direction if we approach $e(A)$ from the other side crossing $C_2 = e(\partial_2 A)$. Curiously enough, Dehn himself [44,45] did not distinguished between left and right twists in the notations (or otherwise), but always described explicitly which twist he deals with.

Let us consider now twist diffeomorphisms about trivial circles. If $e(a)$ bounds a disc D in S, we may assume that D contains $\partial_1 A$ (and not $\partial_2 A$), replacing, if necessary, e by $e \circ i$, where $i: A \to A$ is defined by the formula $i(r, \theta) = (3 - r, 2\pi - \theta)$ (clearly, $e(a) = e \circ i(a)$, but e and $e \circ i$ belong to different isotopy classes). Then we can extend e to an embedding $e': D_2 \to S$ of the disc $D_2 = \{(r, \theta)\colon\ r \leqslant 2\}$ such that $e'(D_{3/2}) = D$, where $D_{3/2} = \{(r, \theta)\colon\ r \leqslant 3/2\}$. We can use e' to transplant the isotopy $\{T'_u : D_2 \to D_2\}_{0 \leqslant u \leqslant 1}$ of D_2 defined by the formula $T'_u(r, \theta) = (r, \theta + \varphi(r) + u(2\pi - \varphi(r)))$ from D_2 to S in the same manner as we transplanted T from A to S. Clearly, the isotopy $\{T'_u\}_{0 \leqslant u \leqslant 1}$ is fixed on the boundary $\partial D_2 = \partial_2 A$ and connects the extension T'_0 of T to D_2 by the identity with $T'_1 = \mathrm{id}_{D_2}$ (note that $(r, \theta) = (r, \theta + 2\pi)$). Hence, our transplanted isotopy connects T_e with the identity diffeomorphism of S and is fixed on ∂S. Therefore, any twist diffeomorphism about a circle bounding a disc in S is isotopic to the identity by an isotopy fixed on ∂S. A similar argument shows that any twist diffeomorphism about a circle C bounding an annulus together with a boundary component of S (or equal to a such component) is isotopic to the identity; this time by an isotopy moving this boundary component.

It follows that, in PMod_S, any Dehn twist about a trivial circle (i.e., a circle bounding a disc or homotopic to a boundary component) is trivial (i.e., equal to 1). But in $\mathcal{M}_S$ only the Dehn twists about circles bounding a disc are trivial and the Dehn twists about circles homotopic to a boundary component are nontrivial (of course, the last statement requires a proof).

Now, we prove some basic properties of Dehn twists. We start with a refinement of the second statement of Corollary 2.7.E.

LEMMA 4.1.A. *The group $\pi_0(\mathrm{Diff}(A \operatorname{fix} \partial A))$ is an infinite cyclic group generated by the isotopy class of the standard twist diffeomorphism T.*

PROOF. Consider the fibration $\mathrm{Diff}(D_2 \operatorname{fix} \partial) \to \mathrm{Emb}(D_1, D_2)$ from Theorem 2.6.A (recall that $D_i = \{(r, \theta)\colon\ r \leqslant i\}$; $i = 1, 2$). Since $\mathrm{Diff}(D_2 \operatorname{fix} \partial)$ is contractible (by Theorem 2.7.C) and since $\mathrm{Emb}(D_1, D_2)$ is weakly homotopy equivalent to the unit tangent bundle of D^2 (by Theorem 2.6.C) and hence to a circle, the homotopy sequence of our fibration implies that $\pi_0(\mathrm{Diff}(A \operatorname{fix} \partial A))$ is isomorphic to $\mathbf{Z}$. Obviously, $\pi_1(\mathrm{Emb}(D_1, D_2))$ is generated by the homotopy class of the loop uniformly rotating D_1 once around its center. To compute the image of this generator in $\pi_0(\mathrm{Diff}(A \operatorname{fix} \partial A))$ we need to extend this rotation to an isotopy of D_2 fixed on ∂D_2. This was done, in fact, in our discussion of twist diffeomorphisms about trivial circles above. The extended isotopy ends in a diffeomorphism equal to the identity on D_1 and to the standard twist diffeomorphism on A. The lemma follows. □

COROLLARY 4.1.B. *Let C be circle on S, and let $H, H' : S \to S$ be two orientation-preserving diffeomorphisms leaving C invariant and preserving the orientation of C. If the results of cutting H and H' along C are isotopic (with free boundary), then, up to isotopy, H differs from H' by a power of a twist diffeomorphism about C.*

PROOF. Changing H and H', if necessary, by an isotopy, we may assume that both H and H' are equal to the identity on C (because $\mathrm{Diff}(C)$ is connected). Obviously, both H and H' preserve the sides of C, and hence H and H' may be assumed to be equal to the identity on some annulus X having C as one of the boundary components. If we remove the interior of X from S, we get a surface S_X diffeomorphic to the result S_C of cutting S along C. Obviously, the diffeomorphisms $H_X, H'_X : S_X \to S_X$ induced by H, H' are isotopic (after a natural identification of S_X with S_C) to the results of cutting H, H' along C. Hence, H_X, H'_X are isotopic. By extending this isotopy to S, we get an isotopy between H and some diffeomorphism H'' which is equal to H' outside X and hence is equal to $H' \circ G$, where G is some diffeomorphism supported[1] in X. Since X is an annulus, an application of the lemma completes the proof. □

LEMMA 4.1.C. *If $f \in \mathrm{Mod}_S$ and γ is the isotopy class of a circle on S, then*

$$f t_\gamma f^{-1} = t_{f(\gamma)}.$$

PROOF. If $e : A \to S$ is an embedding and $F : S \to S$ is a diffeomorphism, then, clearly, $F \circ T_e \circ F^{-1} = T_{F \circ e}$. The lemma follows. □

COROLLARY 4.1.D. *Any two Dehn twists about nonseparating circles on S are conjugate.*

PROOF. It is sufficient to apply Lemma 2.3.A. □

LEMMA 4.1.E. *Let C and D be two circles on S transversely intersecting at one point and let γ, δ, respectively, be their isotopy classes.*

(i) $t_\gamma t_\delta(\gamma) = \delta$.

(ii) $t_\delta t_\gamma^2 t_\delta(\gamma) = \gamma$, *and $t_\delta t_\gamma^2 t_\delta$ can be represented by a diffeomorphism taking C to C and reversing the orientation of C (or, what is the same, interchanging the sides of C).*

[1] The support of a diffeomorphism f is the closure of the set $\{x : f(x) \neq x\}$.

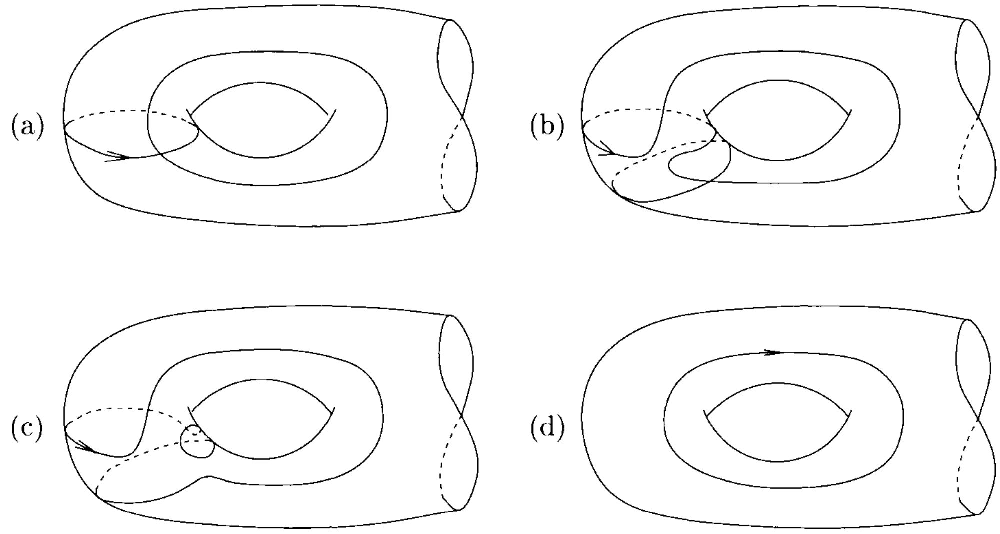


Fig. 4.

PROOF. First, notice that for any two pairs of circles transversely intersecting at one point there is a diffeomorphism of S taking one pair to the other. So, it is sufficient to prove the lemma for one such pair; for example, for the pair C, D in Figure 4(a) (the orientations of circles play a role only in the proof of (ii)).

(i) The image E of C under a twist diffeomorphism about D is illustrated in Figure 4(b), together with a circle isotopic to C. The image of E under a twist diffeomorphism about this last circle is illustrated in Figure 4(c). Clearly, the circle in Figure 4(c) is isotopic to the circle in Figure 4(d), which is nothing other than D. This proves (i).

(ii) In order to prove that $t_\delta t_\gamma^2 t_\delta(\gamma) = \gamma$, it is sufficient to apply (i) twice. In order to prove the remaining part of (ii), one needs to trace the orientations of our circles. The image of C oriented as in Figure 4(a) under the composition of two twist diffeomorphisms from the proof of (i) is illustrated in Figure 4(d). This image is isotopic to D with the orientation given by turning counterclockwise the orientation of C at the (unique) intersection point of C and D. If we apply these two diffeomorphism once more, but in the other order (so we get an oriented representative of $t_\delta t_\gamma^2 t_\delta(\gamma)$), we get C with the orientation given by turning the last orientation of D again counterclockwise. Clearly, this orientation of C is opposite to the original one. This proves (ii). □

I learned Lemma 4.1.E from Birman's survey [17], but the assertion (i) is definitely due to Dehn [45] (cf. [45, §4a])), as are also Lemma 4.1.C and Corollary 4.1.D. Now we prove several fundamental relations between Dehn twists.

LEMMA 4.1.F. *Let C and D be two circles on S and let γ, δ, respectively, be their isotopy classes.*

(i) *If $C \cap D = \emptyset$, then $t_\gamma t_\delta = t_\delta t_\gamma$.*

(ii) *If C and D intersect transversely at exactly one point, then the following* Artin relation *holds*:

$$t_\gamma t_\delta t_\gamma = t_\delta t_\gamma t_\delta.$$

PROOF. (i) If $C \cap D = \emptyset$, then, obviously, t_γ and t_δ can be represented by twist diffeomorphisms with disjoint supports. Such diffeomorphisms obviously commute.

(ii) $t_\gamma t_\delta t_\gamma = t_\delta t_\gamma t_\delta$ is obviously equivalent to $t_\delta t_\gamma t_\delta^{-1} = t_\gamma^{-1} t_\delta t_\gamma$. By Lemma 4.1.C, the latter equation is equivalent to

$$t_{t_\delta(\gamma)} = t_{t_\gamma^{-1}(\delta)}.$$

Hence, it is sufficient to prove that $t_\delta(\gamma) = t_\gamma^{-1}(\delta)$, or, what is the same, $t_\gamma t_\delta(\gamma) = \delta$. But this is the content of Lemma 4.1.E(ii). □

LEMMA 4.1.G. *Let C and D be two circles on S transversely intersecting at one point and let γ, δ, respectively, be their isotopy classes. Let N be a regular neighborhood of the union $C \cup D$ (this means that N cut along $C \cup D$ is an annulus with corners) and let $B = \partial N$, $\beta = \langle B \rangle$. Then*

$$(t_\gamma t_\delta)^6 = t_\beta.$$

PROOF. Instead of N we may consider the standard torus with one hole resulting from the identification of the opposite sides of a square having a small hole around its center. As C and D we may take circles represented by a vertical and a horizontal segment in this square, respectively. Let V and H be the corresponding twist diffeomorphisms.

Consider the horizontal arc h in Figure 5(a). Figures 5(a)–(d) compute $H \circ V(h)$ (up to isotopy). If we apply V to the arc in Figure 5(d), we get the same arc. So, the arc in Figure 5(d) is isotopic to $V \circ H \circ V(h)$. Outside the big circle in Figure 5(e) the arc in Figure 5(d) looks exactly as the arc in Figure 5(a) turned 90°. Such a turn preserves orientation and hence interchanges V and H (more precisely, conjugates one to the other). So, by turning Figures 5(a)–(d), we can find the image of the arc in Figure 5(d) under $H \circ V \circ H$, i.e., we can find $(H \circ V)^3(h) = (H \circ V \circ H) \circ (V \circ H \circ V)(h)$. The result is the arc in Figure 5(f). This arc coincides with h outside the big circle in Figure 5(g). This allows us to find the image of the arc in Figure 5(f) under $(H \circ V)^3$ or, what is the same, $(H \circ V)^6(h)$. Clearly, the result is isotopic to the image of h under a twist diffeomorphism about the boundary (the small circle in our pictures).

Next, we would like to find $(H \circ V)^6(v)$, where v is a vertical arc similar to h; cf. Figure 5(h). Using a 90° turn as above, we see that it is sufficient to find $(V \circ H)^6(h)$. Note that the first H in $(V \circ H)^6$ acts on h trivially, so $(V \circ H)^6(h) = (V \circ H)^5 \circ V(h) = V \circ (H \circ V)^5(h)$. But $V \circ (H \circ V)^5(h) = (H \circ V)^6(h)$, because in the computation of $(H \circ V)^6(h)$ the last H acts trivially (like the last V in the computation of $V \circ H \circ V(h)$). Hence, $(V \circ H)^6(h) = (H \circ V)^6(h)$, and hence $(V \circ H)^6(h)$ is equal to the image of h under a twist diffeomorphism about the boundary. It follows that the image $(H \circ V)^6(v)$ is equal to the image of v under a twist diffeomorphism about the boundary.

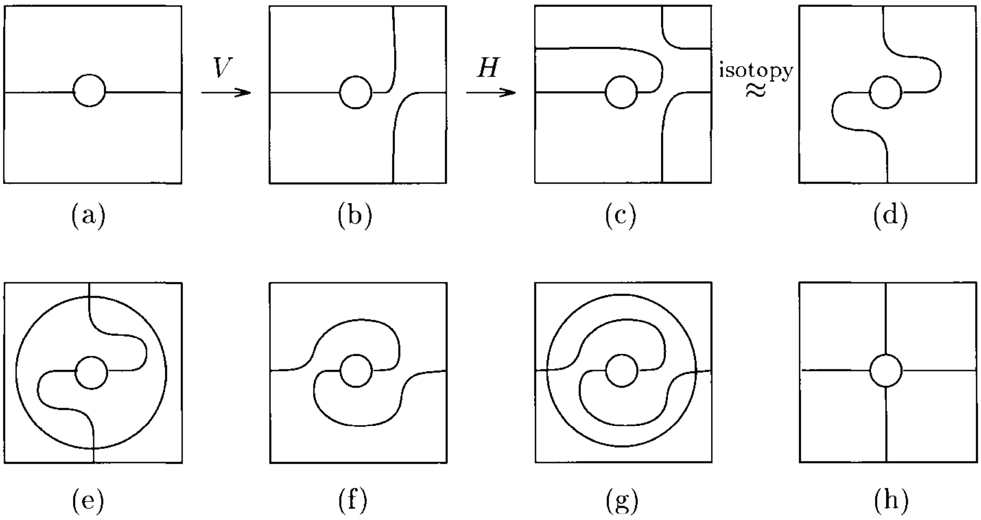


Fig. 5.

It remains to notice that if we cut our torus with a hole along h and v, we get a disc (with corners), and hence a diffeomorphism of our torus (fixed on the boundary) is determined, up to an isotopy fixed on the boundary, by its action on h and v (recall that $\mathrm{Diff}(D \operatorname{fix} \partial D)$ is connected for any disc D). □

This lemma is also due (in a slightly different form) to Dehn [45] (cf. [45, §6c]).

LEMMA 4.1.H. *Let S_0^4 be a sphere with four holes. Let the boundary components of S_0^4 be $C_0, \ldots, C_3$ and for $1 \leqslant i < j \leqslant 3$, let C_{ij} denote a circle encircling C_i and C_j as in Figure 6. Suppose that S_0^4 is embedded in S. Let $t_i \in \mathcal{M}_S$ be the Dehn twist about C_i,*

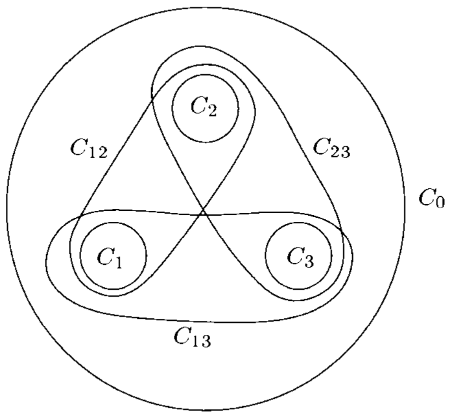


Fig. 6.

$0 \leqslant i \leqslant 3$, *and let* $t_{ij} \in \mathcal{M}_S$ *be the Dehn twist about* C_{ij}, $1 \leqslant i < j \leqslant 3$. *Then the following* lantern relation *holds in* $\mathcal{M}_S$ *(and hence in* Mod_S*)*:

$$t_0 t_1 t_2 t_3 = t_{12} t_{13} t_{23}.$$

PROOF. Connect C_0 with C_1, C_2, C_3 by three disjoint arcs I_1, I_2, I_3. Clearly, if we cut S_0^4 along these three arcs, we get a disc (with corners). Hence (as in the proof of Lemma 4.1.G), a diffeomorphism of S_0^4 fixed on ∂S_0^4 is determined up to an isotopy fixed on ∂S_0^4 by its action on I_1, I_2, I_3. Therefore, in order to prove our relation it is sufficient to compute the action of some representatives of both sides on the arcs I_1, I_2, I_3. Of course, we will take compositions of twist diffeomorphisms as our representatives, and then their action on the arcs can be computed in a straightforward manner, as in the proof of Lemma 4.1.E. One just needs to draw several pictures, and we leave this task to the reader. □

The relation of Lemma 4.1.H was discovered by Dehn [45] (cf. [45, §7g) 1)]) and much later rediscovered and popularized under the name *lantern relation* by Johnson [113].

Finally, we will prove a lemma borrowed from [17] (cf. [17], the discussion after Corollary 3.3). It is crucial for the induction step in some arguments using the induction on the number of boundary components. Cf. Section 4.2.

LEMMA 4.1.I. *Suppose that* $\partial S \neq \emptyset$. *Let* R *be the result of gluing a disc* D *to one of the boundary components of* S. *Let* $\beta \in \pi_1(R, x)$ *be an element represented by an* embedded *loop* b. *Without any loss of generality we may assume that the base point* x *is contained in the interior of* D *and the loop* b *intersects* D *in a single arc containing* x. *Then we can choose an annulus* B *containing* D *and* b *in its interior and having boundary components isotopic to* b. *Let* C_1 *be the component of* ∂B *lying at the left side of* b, *and let* C_2 *be the component of* ∂B *lying at the right side of* b, *where the meaning of* left *and* right *is determined by the direction of the loop* b *and the orientation of* S. *Then*

$$j(\beta) = t_{\gamma_1} t_{\gamma_2}^{-1},$$

where $j : \pi_1(R, x) \to \mathrm{PMod}_S$ *is the homomorphism from* 2.8, *and* $\gamma_1 = \langle C_1 \rangle$, $\gamma_2 = \langle C_2 \rangle$.

PROOF. $j(\beta)$ can be obtained in the following manner (cf. 2.8). Consider the loop b as an isotopy of the base point x and extend this isotopy to an isotopy $\{F_t : R \to R\}_{0 \leqslant t \leqslant 1}$, $F_0 = \mathrm{id}_R$. We may assume that the ending diffeomorphism F_1 of this isotopy is fixed on D. Then it induces a diffeomorphism $F : S \to S$, and $j(\beta)$ is equal to the isotopy class of F. Obviously, the isotopy $\{F_t\}_{0 \leqslant t \leqslant 1}$ may be chosen to be fixed outside of the annulus B. Moreover, it can be chosen is such a way that F_1 is equal to the composition of a left twist diffeomorphism supported in an annulus (contained in B) with the boundary C_1 and a right twist diffeomorphism supported in an annulus (disjoint from the first one and also contained in B) with the boundary C_2. One can see this by drawing a couple of simple pictures or, alternatively, by defining $\{F_t\}_{0 \leqslant t \leqslant 1}$ by explicit formulas (compare our discussion of twist diffeomorphisms about trivial circles). We leave this as an exercise to the reader (the pictures can be found in [101, Section (6.1)]). Clearly, this description of F_1 implies the lemma. □

4.2. *The Dehn–Lickorish theorem*

The main result of this section is Theorem 4.2.D, known as the Dehn–Lickorish theorem. It provides an explicit finite set of Dehn twists generating Mod_S (for closed S). We also discuss some corollaries of this result. Let us start with a couple of simple lemmas.

LEMMA 4.2.A. *The Dehn twists about the longitude and the meridian of a closed torus S generate* Mod_S.

PROOF. By the discussion after Theorem 2.9.A, Mod_S is isomorphic to $\mathrm{SL}_2(\mathbf{Z})$. This isomorphism, of course, depends on the choice of an isomorphism $\pi_1(S) \to \mathbf{Z}^2$. If the latter isomorphism takes (the elements represented by) the longitude and the meridian into the standard basis of $\mathbf{Z}^2$, then, as it is easy to see, the Dehn twists about the longitude and the meridian correspond to matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

On the other hand, it is well known that these matrices generate $\mathrm{SL}_2(\mathbf{Z})$. The lemma follows. □

LEMMA 4.2.B. *Suppose that S is a closed surface of genus $\geqslant 2$. Let γ, γ' be two vertices of $C(S)$ represented by nonseparating circles on S. Then there is a sequence $\gamma = \gamma_1, \gamma_2, \dots, \gamma_n = \gamma'$ of vertices of $C(S)$ such that any two consecutive vertices γ_i, γ_{i+1} are connected by an edge in $C(S)$ and, moreover, if they are represented by two disjoint circles C_i, C_{i+1}, then the union $C_i \cup C_{i+1}$ does not separate S.*

PROOF. Since $C(S)$ is connected, there is a sequence $\gamma = \delta_1, \delta_2, \dots, \delta_m = \gamma'$ of vertices of $C(S)$ such that any two consecutive vertices δ_i, δ_{i+1} are connected by an edge in $C(S)$. It may happen that some δ_i is the isotopy class of a circle D_i separating S into two parts S_1, S_2. In this case δ_{i-1}, δ_{i+1} can be represented by two circles D_{i-1}, D_{i+1}, respectively, disjoint from D_i. If D_{i-1} and D_{i+1} are contained in different parts of S, then $D_{i-1} \cap D_{i+1} = \emptyset$ and we can delete δ_i from our sequence. Suppose now that both D_{i-1} and D_{i+1} are contained in the same part of S, say in S_1. Since S is closed and D_i is nontrivial, the genus of S_2 is $\geqslant 1$ and hence S_2 contains some nonseparating circle D_i'. In this case we replace δ_i by $\langle D_i' \rangle$. After repeating this procedure several times, we will get a new sequence $\gamma = \delta_1, \delta_2, \dots, \delta_m = \gamma'$ such that all δ_i are the isotopy classes of nonseparating circles (and any two consecutive vertices δ_i, δ_{i+1} are connected by an edge).

Now, let $1 \leqslant i \leqslant m-1$ and let D_i, D_{i+1} be two disjoint circles representing δ_i, δ_{i+1} respectively. Suppose that $D_i \cup D_{i+1}$ separates S into two parts S_1, S_2. Since S is closed, both these parts have genus $\geqslant 1$ (we assume that $\delta_i \neq \delta_{i+1}$). It follows that S_1 (as also S_2) contains some nonseparating circle D_i'. Clearly, both $D_i \cup D_i'$ and $D_i' \cup D_{i+1}$ do not separate S.

Let us replace the part δ_i, δ_{i+1} of our sequence by $\delta_i = \langle D_i \rangle$, $\delta_i' = \langle D_i' \rangle$, $\delta_{i+1} = \langle D_{i+1} \rangle$. After repeating this procedure several times, we will get the required sequence. □

THEOREM 4.2.C. *If S is a compact surface of genus $\geqslant 1$, then* PMod_S *is generated by Dehn twists about nonseparating circles.*

PROOF. We use the double induction on the genus and the number of boundary components. The induction starts with the closed torus. In this case the theorem follows from Lemma 4.2.A.

Suppose that $\partial S \neq \emptyset$. Let R be the result of gluing a disc to one of the boundary components of S. By the inductive assumption, PMod_R is generated by twists about nonseparating circles. Obviously, every circle C on R is isotopic to a circle C' contained in S, and $t_C = \rho(t_{C'})$, where $\rho : \mathrm{PMod}_S \to \mathrm{PMod}_R$ is the homomorphism from 2.8. Moreover, if C is nonseparating, then C' is also nonseparating. It follows that in order to prove the theorem for S it is sufficient to prove that $\ker \rho$ is generated by Dehn twists about nonseparating circles. The latter follows from Lemma 4.1.I and the fact that $\pi_1(R)$ is generated by the homotopy classes of (embedded) nonseparating loops. This completes the induction step of the induction on the number of boundary components.

Next, suppose that S is closed. Let $f \in \mathrm{Mod}_S$. Let us choose a nonseparating circle C on S. Consider two vertices $\gamma = \langle C \rangle$, $\gamma' = f(\gamma)$ of $C(S)$ and connect them by a sequence $\gamma = \gamma_1, \gamma_2, \ldots, \gamma_m = \gamma'$ as in Lemma 4.2.B. For every $i = 1, 2, \ldots, m-1$ we can represent γ_i, γ_{i+1} by disjoint circles C_i, C_{i+1}. Since $C_i \cup C_{i+1}$ does not separate S, we can choose a circle D_i transversely intersecting each of C_i and C_{i+1} at exactly one point. Let $\delta_i = \langle D_i \rangle$. By Lemma 4.1.E(i), $t_{\gamma_{i+1}} t_{\delta_i}(\gamma_{i+1}) = \delta_i$ and $t_{\delta_i} t_{\gamma_i}(\delta_i) = \gamma_i$. Hence, $g(\gamma') = \gamma$ and $g \circ f(\gamma) = \gamma$, where g is the product of elements $t_{\delta_i} t_{\gamma_i} t_{\gamma_{i+1}} t_{\delta_i}$, $i = 1, 2, \ldots, m-1$. Since $g \circ f(\gamma) = \gamma$, we can represent $g \circ f$ by a diffeomorphism $H : S \to S$ preserving C. If H interchanges the sides of C, we replace g by $t_{\delta_1} t_{\gamma_1}^2 t_{\delta_1} g$. Lemma 4.1.E(ii) implies that for the new g the element $g \circ f$ can be represented by a diffeomorphism $H : S \to S$ preserving C and preserving the sides of C. Now, let us cut S and H along C. We will get a diffeomorphism $H_C : S_C \to S_C$ preserving both boundary components of S_C. Hence, the isotopy class h_C of H_C belongs to PMod_{S_C}. Since the genus of S_C is less than the genus of S, h_C can be presented, by the inductive assumption, as a product of Dehn twists about nonseparating circles on S_C. Every circle on S_C is at the same time a circle on S, and hence these Dehn twists in PMod_{S_C} naturally define Dehn twists in PMod_S. The product of these Dehn twists in PMod_S can be represented by a diffeomorphism $H' : S \to S$ preserving C and isotopic to H_C after cutting along C. It follows that H differs from H' (up to isotopy) by some power of a Dehn twist about C (cf. Corollary 4.1.B). Hence, $g \circ f$ is equal to a product of Dehn twists about nonseparating circles. Since g is also such a product, by its construction, we conclude that f is a product of Dehn twists about nonseparating circles. This completes the induction step of the induction on genus, and hence the proof. □

This theorem is due to Dehn [45]. His work was forgotten for a while, and the theorem was rediscovered by Lickorish [132]. To a certain extent, our proof follows the proof of Birman [17] (cf. [17, Theorem 4.1]), who simplified the proof of Lickorish. The use of the connectedness of $C(S)$ leads to further radical simplifications, and is, in fact, closer in

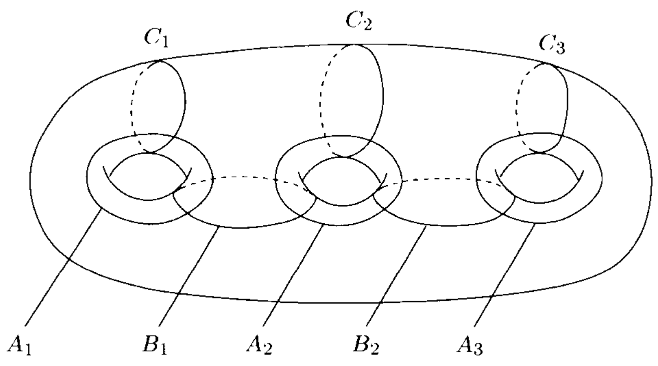


Fig. 7.

spirit to Dehn's original approach than to the one of Lickorish. The main tool of Dehn [45] was the action of Mod_S on the set of (the isotopy classes of) systems of circles. What was lacking in [45] (and is crucial to our approach) is an appropriate structure on this set, namely, the structure of a simplicial complex introduced much later by Harvey [87].

Now we are ready to prove the Dehn–Lickorish theorem.

THEOREM 4.2.D. *If S is a closed surface of genus g, then* Mod_S *is generated by Dehn twists about the $3g-1$ circles $A_1, \ldots, A_g, B_1, \ldots, B_{g-1}, C_1, \ldots, C_g$ presented in Figure 7.*

PROOF. By Lemma 4.2.A the theorem holds for $g = 1$. Using induction on g, we may assume that the theorem holds for all surfaces of genus $\leqslant g-1$.

Let us denote by $\alpha_1, \ldots, \alpha_g, \beta_1, \ldots, \beta_{g-1}, \gamma_1, \ldots, \gamma_g$ the isotopy classes of the circles $A_1, \ldots, A_g, B_1, \ldots, B_{g-1}, C_1, \ldots, C_g$ respectively. Let us denote by G_S the subgroup of Mod_S generated by Dehn twists about these $3g-1$ circles. So, we have to prove that $G_S = \mathrm{Mod}_S$. In view of Theorem 4.2.C it is sufficient to prove that G_S contains all Dehn twists about nonseparating circles. By its definition, the subgroup G_S contains some such twists. Hence, by Lemma 4.1.C it is sufficient to prove that G_S acts transitively on the set of isotopy classes of nonseparating circles. Moreover, since C_1 is nonseparating, it is sufficient to prove that for every nonseparating circle C' there is an element $f \in G_S$ such that $f(\gamma_1) = \gamma'$, where $\gamma' = \langle C' \rangle$.

Now we claim that it is sufficient to prove the last statement only for circles C' such that $C_1 \cap C' = \emptyset$ and $C_1 \cup C'$ does not separate S. In fact, suppose that our statement is proved for such special circles C' and consider an arbitrary nonseparating circle C'. Let us connect γ_1 with $\gamma' = \langle C' \rangle$ by a sequence $\gamma_1 = \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1n} = \gamma'$ as in Lemma 4.2.B. By our assumption, there is $g \in G_S$ such that $g(\gamma_1) = g(\gamma_{11}) = \gamma_{12}$. Note that $\gamma_1 = \gamma_{11}$ is connected with $g^{-1}(\gamma')$ by a shorter sequence $\gamma_1 = \gamma_{11} = g^{-1}(\gamma_{12}), g^{-1}(\gamma_{13}), \ldots, g^{-1}(\gamma_{1n}) = g^{-1}(\gamma')$ with the properties of Lemma 4.2.B. Using induction, we may assume that $g^{-1}(\gamma') = h(\gamma_1)$ for some $h \in G_S$. Then $g \circ h(\gamma_1) = \gamma'$ and $g \circ h \in G_S$. This proves our claim.

So, let C' be a circle such $C_1 \cap C' = \emptyset$ and $C_1 \cup C'$ does not separate S, and let $\gamma' = \langle C' \rangle$. By Lemma 4.1.E(i) we have $t_{\gamma_1} t_{\alpha_1}(\gamma_1) = \alpha_1$, $t_{\alpha_1} t_{\beta_1}(\alpha_1) = \beta_1$, $t_{\beta_1} t_{\alpha_2}(\beta_1) = \alpha_2$, and

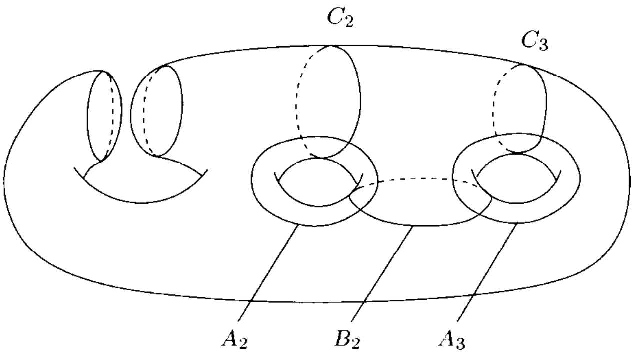


Fig. 8.

$t_{\alpha_2} t_{\gamma_2}(\alpha_2) = \gamma_2$. It follows that $g(\gamma_1) = \gamma_2$ for some $g \in G_S$ (namely, for the product of the above eight Dehn twists). Therefore, it is sufficient to prove that $f(\gamma_2) = \gamma'$ for some $f \in G_S$. Note that both C_2 and C' are disjoint from C_1 and both $C_1 \cup C_2$ and $C_1 \cup C'$ do not separate S.

Let R be the result of cutting S along C_1, and let Q be the result of gluing two discs D_1, D_2 to the two boundary components of R. Clearly, both C_2 and C' are contained in Q (and even in R) and do not separate Q. Consider the subgroup G_Q of Mod_Q generated by Dehn twists about the circles $A_2, \dots, A_g, B_2, \dots, B_{g-1}, C_2, \dots, C_g$ (clearly, all these circles are contained in Q). Cf. Figure 8. Since the genus of Q is equal to $g - 1$, we have $G_Q = \mathrm{Mod}_Q$ by the inductive assumption.

Since $G_Q = \mathrm{Mod}_Q$, we can find a diffeomorphism $F : Q \to Q$ such that $F(C_2)$ is isotopic to C' in Q and the isotopy class of F belongs to G_Q. Since G_Q is generated by the Dehn twists about the circles $A_2, \dots, A_g, B_2, \dots, B_{g-1}, C_2, \dots, C_g$, we may assume that F is equal to a composition of twist diffeomorphisms supported in some annuli with boundary components isotopic to these circles. Clearly, we may assume these annuli to be disjoint from the discs D_1, D_2 added to R. In particular, we may assume that F is fixed on D_1 and D_2. Such a diffeomorphism F can be obtained by cutting a (unique) diffeomorphism $F^{\sim} : S \to S$ fixed on C_1 along C_1 and gluing in the identity diffeomorphisms id_{D_1}, id_{D_2}. Obviously, $F^{\sim}$ is equal to a composition of twist diffeomorphisms supported in the same annuli as F. In particular, its isotopy class $f^{\sim}$ belongs to G_S.

If $F(C_2)$ is not only isotopic, but equal to C', then $F^{\sim}(C_2)$ is also equal to C', and hence $f^{\sim}(\gamma_2) = \gamma'$. This completes the proof in the case $F(C_2) = C'$, since $f^{\sim} \in G_Q$. In the rest of the proof we deal with the complications arising from the fact that in general we only know that $F(C_2)$ is isotopic to C' in Q.

Let $\{K_t : Q \to Q\}_{0 \leqslant t \leqslant 1}$ be an isotopy moving $F(C_2)$ into C', so that $K_0 = \mathrm{id}_Q$, $K_1(F(C_2)) = C'$. During this isotopy, the discs D_1, D_2 may be moved. Note that D_1 and D_2 are disjoint from $F(C_2)$ (because F is fixed on D_1 and D_2), and hence $K_1(D_1)$ and $K_1(D_2)$ are disjoint from C'. Since C' does not separate Q, we can move the discs $K_1(D_1)$, $K_1(D_2)$ back to their original position D_1, D_2 without moving C'. More precisely, there is an isotopy $\{L_t : Q \to Q\}_{0 \leqslant t \leqslant 1}$ such that $L_0 = \mathrm{id}_Q$, L_t is fixed on C' for all t, and $L_1 \circ K_1 \mid D_1 = \mathrm{id}_{D_1}$, $L_1 \circ K_1 \mid D_2 = \mathrm{id}_{D_2}$. Let $H = L_1 \circ K_1$. Clearly, $H(F(C_2)) = C'$,

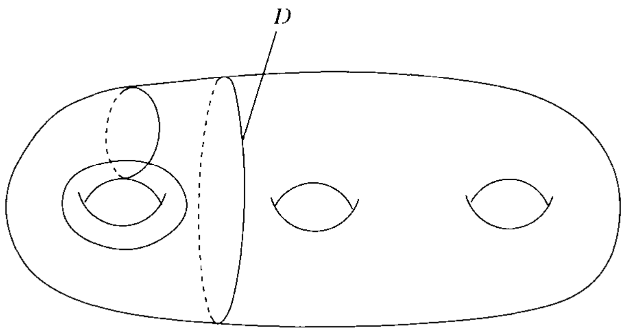


Fig. 9.

H is isotopic to id_Q and $H \mid D_1 = \mathrm{id}_{D_1}$, $H \mid D_2 = \mathrm{id}_{D_2}$. Like F, the diffeomorphism H can be obtained by cutting a (unique) diffeomorphism $H^{\sim} : S \to S$ fixed on C_1 along C_1 and gluing in the identity diffeomorphisms id_{D_1}, id_{D_2}. Let $h^{\sim}$ be the isotopy class of $H^{\sim}$.

Since $H \circ F(C_2) = C'$, we have $H^{\sim} \circ F^{\sim}(C_2) = C'$ and hence $h^{\sim} \circ f^{\sim}(\gamma_2) = \gamma'$. Recall that $f^{\sim} \in G_S$ and hence $h^{\sim} \circ f^{\sim} \in G_S$ if $h^{\sim} \in G_S$. Therefore, in order to complete the proof it is sufficient to show that $h^{\sim} \in G_S$. In view of the properties of H stated in the last paragraph, it is sufficient, in fact, to prove the following claim.

CLAIM. *Let $H^{\sim} : S \to S$ be a diffeomorphism fixed on C_1, preserving the sides of C_1 and such that the result H of cutting $H^{\sim}$ along C_1 and gluing in the identity diffeomorphisms id_{D_1}, id_{D_2} is isotopic to id_Q. Then the isotopy class $h^{\sim}$ of $H^{\sim}$ belongs to G_S.*

Let us now prove this claim. First, by Lemma 4.1.E(ii) some representative $I^{\sim}$ of the isotopy class $i^{\sim} = t_{\alpha_1} t_{\gamma_1}^2 t_{\alpha_1}$ maps C_1 to C_1 and interchanges the sides of C_1. Let $I' : R \to R$ be the result of cutting $I^{\sim}$ along C_1 and let $i' \in \mathrm{Mod}_R$ be its isotopy class. Next, by Lemma 4.1.G, $t^{\sim} = (t_{\gamma_1} t_{\alpha_1})^6$ is equal to t_{δ_0}, where δ_0 is the isotopy class of the circle D_0 in Figure 9. In particular, $t^{\sim}$ can be represented by a twist diffeomorphism $T^{\sim}$ about D_0. We may assume that $T^{\sim}$ is supported in an annulus disjoint from C_1. Let $T' : R \to R$ be the result of cutting $T^{\sim}$ along C_1 and let $t' \in \mathrm{Mod}_R$ be its isotopy class. Let G be the subgroup of Mod_R generated by i', t' and the Dehn twists about $A_2, \dots, A_g, B_2, \dots, B_{g-1}, C_2, \dots, C_g$.

Let $H_C^{\sim} : R \to R$ be the result of cutting H along C (recall that $R = S_C$) and let $h_C^{\sim} \in \mathrm{Mod}_R$ be its isotopy class. Suppose that $h_C^{\sim} \in G$, i.e., that $h_C^{\sim}$ can be represented as a product of i', t' and the Dehn twists about $A_2, \dots, A_g, B_2, \dots, B_{g-1}, C_2, \dots, C_g$ on R. Then $h^{\sim}$ differs from the corresponding product of $i^{\sim}$, $t^{\sim}$ and the Dehn twists about $A_2, \dots, A_g, B_2, \dots, B_{g-1}, C_2, \dots, C_g$ (considered now on S) by some power of the Dehn twist t_{γ_1} about C_1. Since $i^{\sim} = t_{\alpha_1} t_{\gamma_1}^2 t_{\alpha_1}$ and $t^{\sim} = (t_{\gamma_1} t_{\alpha_1})^6$ belong to G_S, it follows that in this case $h^{\sim} \in G_S$. Therefore, it is sufficient to prove that $h_C^{\sim} \in G$. By the assumption, $h_C^{\sim}$ belongs to the kernel of the natural map $\rho : \mathrm{PMod}_R \to \mathrm{PMod}_Q$ from 2.8. So, it is sufficient to prove that $\ker \rho \subset G$.

The kernel $\ker \rho$ is equal to the image of the canonical homomorphism $j : PB_2(Q) \to \mathrm{Mod}_R$ from 2.8, and hence is generated by the images of the fundamental groups $\pi_1(Q_1)$, $\pi_1(Q_2)$, where Q_i is the result of gluing only one disc D_i to R (and $i = 1, 2$).

(a)

(b)

(c)

(d)

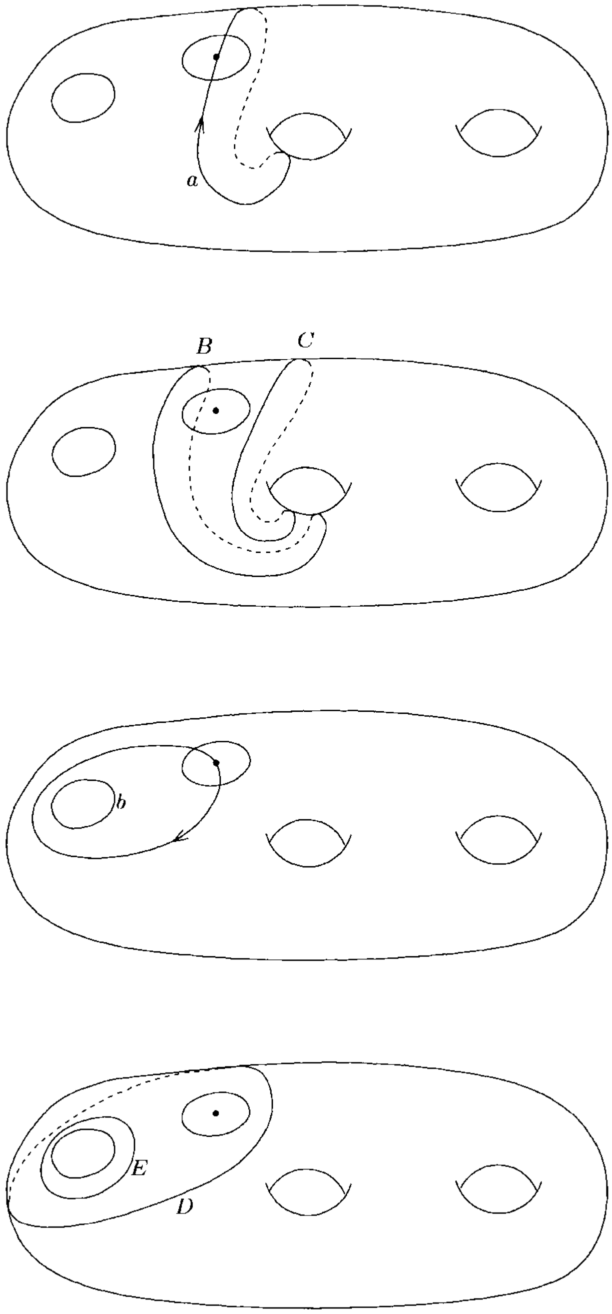


Fig. 10.

Consider first the element a of $\pi_1(Q_2)$ represented by the loop in Figure 10(a). In view of Lemma 4.1.I, its image in Mod_R is equal to $t_\beta t_\gamma^{-1}$, where β, γ are the isotopy classes of circles B, C from Figure 10(b). Clearly, B, C are isotopic in R to B_1, C_2 respectively, and hence this image is equal to $t_{\beta_1} t_{\gamma_2}^{-1} \in G$. Let us consider next the element b of $\pi_1(Q_2)$ represented by the loop in Figure 10(c) (a loop going round the hole). In view of Lemma 4.1.I, its image in Mod_R is equal to $t_\delta t_\varepsilon^{-1}$, where δ, ε are the isotopy classes of circles D, E from Figure 10(d). Clearly, D is isotopic to D_0 and E is a trivial circle (homotopic to a boundary component). It follows that $t_\delta = t_{\delta_0}$, $t_\varepsilon = 1$, and hence $t_\delta t_\varepsilon^{-1} = t_{\delta_0} = t^{\sim}$. Therefore, the image of b also belongs to G.

Recall that $fj(\sigma)f^{-1} = j(f_*(\sigma))$ for any $f \in \mathrm{Mod}_R$, $\sigma \in B_2(Q)$ (cf. 2.8). It follows that $j(f_*(\sigma)) \in G$ if $f \in G$, $j(\sigma) \in G$. In particular, if the image of some element $c \in \pi_1(Q_2)$ belongs to G and $f \in G$, then the image of $f_*(c)$ also belongs to G (if f is not in PMod_R, i.e., if f interchanges the holes, then $f_*(c) \in \pi_1(Q_1)$, but actually we do not need this case). Now, Figures 11(a)–(g) shows how to produce new elements of $\pi_1(Q_2)$ from the element a by applying consecutively various Dehn twists from G to a. Clearly, together with b these elements generate $\pi_1(Q_2)$. It follows that the image of $\pi_1(Q_2)$ in Mod_R is contained in G.

It remains to show the same for the image of $\pi_1(Q_1)$. This can be done by similar pictures. But it is easier to note that conjugation by the element i', interchanging the holes of R, maps the image of $\pi_1(Q_2)$ into the image of $\pi_1(Q_1)$. It follows that (since $i' \in G$) the image of $\pi_1(Q_1)$ is also contained in G. And since the images of $\pi_1(Q_1)$, $\pi_1(Q_2)$ generate $\ker \rho$ by Corollary 2.8.B and Theorem 2.8.C, we conclude that $\ker \rho \subset G$. As we already said, this completes the proof of the claim, and hence of the theorem. □

Dehn [45] was the first to prove that the groups Mod_S are finitely generated and to provide an explicit finite set of generators for them. His set of generators consisted of $2g(g-1)$ Dehn twists for a closed surface S of genus $g \geqslant 3$ and of 5 Dehn twists for a closed surface of genus 2. Much later, and independent of Dehn's work, Theorem 4.2.D was proved by Lickorish [133,137]. An exposition of the work of Lickorish was given by Birman [16]; see [16, Chapter 4]. Again, to a certain extent, our proof follows the simplified proof of Birman [17] (cf. [17, Theorem 4.7 and Corollary 4.8]), and the use of $C(S)$ leads to further radical simplifications and is closer in spirit to Dehn's original approach. (We also added some details corresponding to the proof of Corollary 4.8 from [17], omitted in [17].)

COROLLARY 4.2.E. Mod_S *is finitely generated for any compact (orientable) surface* S.

PROOF. Since PMod_S is of finite index in Mod_S, it is sufficient to prove that the groups PMod_S are finitely generated. Let us glue discs to the boundary components of S one at a time. Using the second exact sequence of Theorem 2.8.C with $n = 1$ (recall that $PB_1(Q) = \pi_1(Q)$ for any surface Q), the fact that the fundamental groups of compact surfaces are finitely generated, and the induction on the number of boundary components, we can reduce our assertion to the case of closed surfaces S. But for closed S, Theorem 4.2.D provides an explicit finite system of generators. □

COROLLARY 4.2.F. *If S is a closed surface of genus g, then* Mod_S *is generated by Dehn twists about the* $2g+1$ *circles* $A_1, \ldots, A_g, B_1, \ldots, B_{g-1}, C_1, C_2$ *presented in Figure* 7.

PROOF. This result is due to Humphries [93], and we refer to [93] for a sequence of pictures showing how to transform C_i into a circle isotopic to C_{i+2} by a sequence of 16

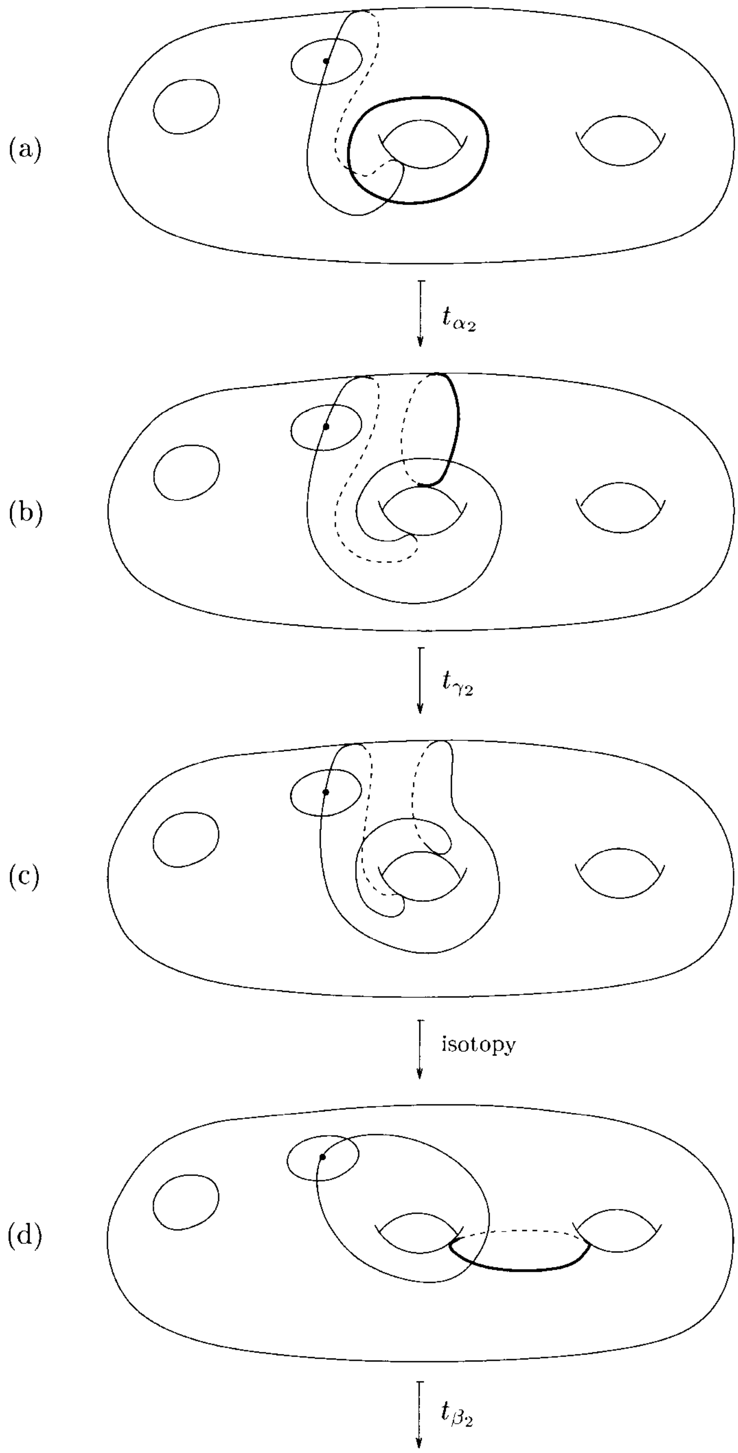


Fig. 11(a)–(d).

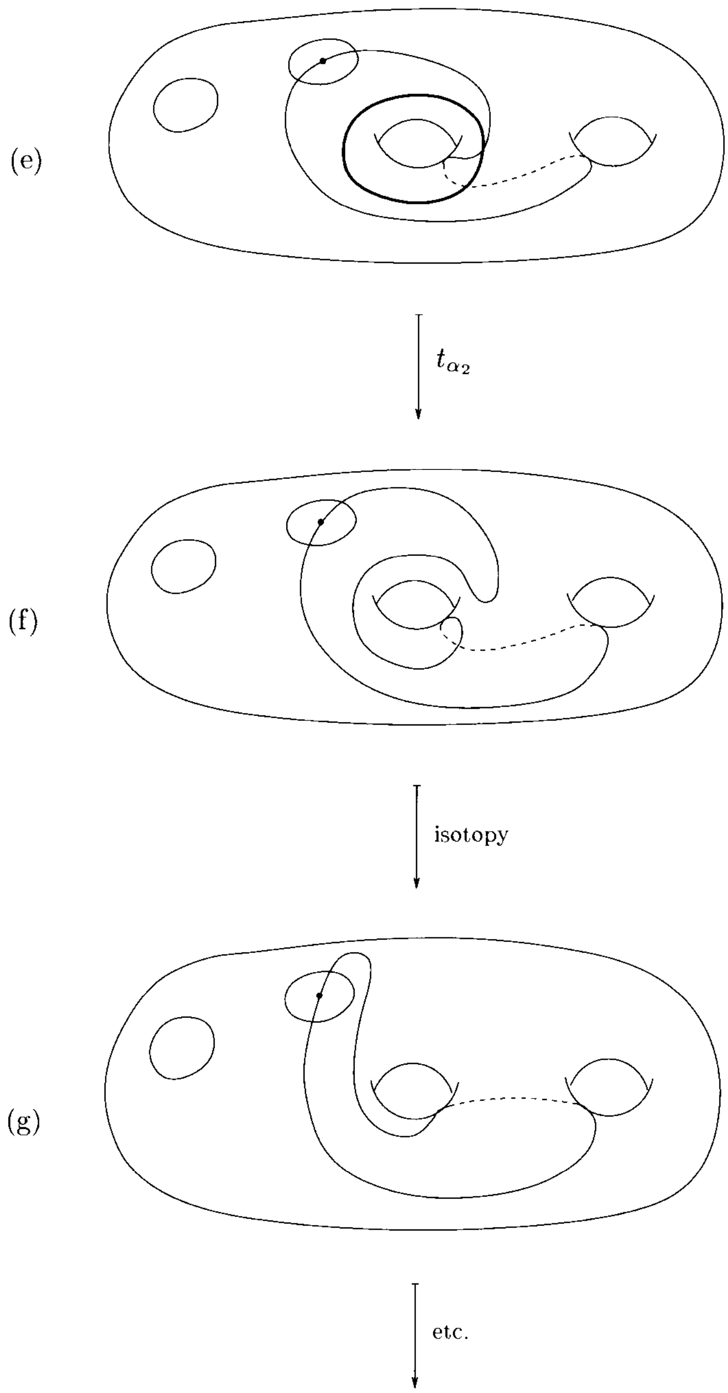


Fig. 11(e)–(g).

twist diffeomorphisms about $A_i, B_i, C_{i+1}, A_{i+1}$, and B_{i+1}. In view of Lemma 4.1.C, this implies that one can express the Dehn twist about C_{i+2} in terms of Dehn twists about $C_i, A_i, B_i, C_{i+1}, A_{i+1}$, and B_{i+1}. It follows that we can consecutively eliminate Dehn twists about $C_g, C_{g-1}, \ldots, C_3$ from our list of generators. □

This result is complemented by a theorem of Humphries [93] to the effect that $2g + 1$ is actually the minimal number of Dehn twist generators of Mod_S if S is a closed surface of genus g. If we do not require the generators to be Dehn twists, then one needs, in fact, only two generators. Namely, the following theorem holds.

THEOREM 4.2.G. *If S is a closed surface of genus g, then* Mod_S *is generated by the two elements*

$$t_{\gamma_1} t_{\alpha_1} t_{\beta_1} t_{\alpha_2} t_{\beta_2} \ldots t_{\beta_{g-1}} t_{\alpha_g},$$
$$t_{\gamma_{g-1}} t_{\gamma_g}^{-1},$$

where $\alpha_1, \beta_1, \ldots, \alpha_g, \gamma_1, \gamma_{g-1}, \gamma_g$ are, respectively, the isotopy classes of the circles $A_1, B_1 \ldots, A_g, C_1, C_{g-1}, C_g$ presented in Figure 7.

This nice result is due to Wajnryb [231]. He deduced it from the above Corollary 4.2.F by some ingenious computations. His proof also covers the case of surfaces with one boundary component (if we made a hole in the surface in Figure 7 at the right end of it, then the Dehn twists about the same circles may serve as generators). The point of this theorem is in providing the *minimal* number of generators (clearly, Mod_S is not cyclic, and hence cannot be generated by one element!). As Wajnryb explains in [231], it is quite easy to find a generating set consisting of three elements, and already Lickorish [134] noticed that four elements generate Mod_S. While the property of being generated by two elements has no immediate important consequences, it traditionally attracts the attention of group theorists.

4.3. *Finite presentations*

After we have proved that the groups Mod_S are finitely generated, it is only natural to ask if they are finitely presented (and to look for a finite presentation, if they are). In the spirit of our proof of the Dehn–Lickorish theorem (i.e., Theorem 4.2.D), we will base our approach to this question upon the action of Mod_S on $C(S)$. Such an approach is made possible by the following general theorem.

THEOREM 4.3.A. *Suppose that a group G acts on a CW-complex X permuting its cells and suppose that:*

(i) *X is simply connected (in particular, X is connected);*
(ii) *the isotropy group of every vertex of X is finitely presented;*
(iii) *the isotropy group of every edge of X is finitely generated;*
(iv) *the number of orbits of cells of dimension $\leqslant 2$ is finite.*

Then G is finitely presented.

An elegant proof of this theorem, based on the Bass–Serre theory [214], is given by Brown [27]; cf. [27, Theorem 4]. But it can also be easily proved directly. In the special case when G acts transitively on the set of vertices, such a proof is contained in the report of Laudenbach [130] about the paper of Hatcher and Thurston [91]; cf. [130, §2]. Implicitly this special case was proved and used already by Hatcher and Thurston [91]. Laudenbach's proof can be easily extended to the general case. Brown [27] gives a long list of precursors of Theorem 4.3.A, but the papers [91,130] apparently were unknown to him. It is worth adding to his list also the results of Behr [8, Section 1], and of Kozsul [129, Chapter III, §2]. While Behr [8] states his sufficient condition for the finite presentability in terms of actions on discrete metric spaces, Koszul [129] reworked it in the language of actions on graphs, and obtained, in fact, a necessary and sufficient condition. Actually, as it is clear from his definition[2] of the homotopy of loops, Koszul [129] works not with graphs, but with flag complexes (cf. 3.1). Compared with [129], the main novelty of Theorem 4.3.A lies in the fact that locally infinite complexes are admitted. According to Harvey [87], the fact that the complexes of curves are not locally finite also was an (at least a psychological) obstacle to overcome.

Note that all proofs of Theorem 4.3.A allow, in principle, to construct a finite presentation of G starting from finite presentations of the isotropy groups of vertices, finite sets of generators of the isotropy groups of edges, the combinatorial structure of X, and, of course, the action of G on X. This leads usually to a fairly complicated presentation, which one may try to simplify. For Mod_S, such an approach was successfully realized by Wajnryb in his work [230] discussed below.

The following two lemmas admit direct and elementary proofs. But we prefer to deduce them from Theorem 4.3.A and the well known fact that for any finitely presented group there is a finite CW-complex having this group as its fundamental group. Our proof of the first lemma was suggested by Koszul [129, §III.2, Remark 2].

LEMMA 4.3.B. *Let $1 \to H \to G \to F \to 1$ be an exact sequence of groups. If H and F are finitely presented, then G is also finitely presented.*

PROOF. Let Y be a finite CW-complex with $\pi_1(Y) = F$, and let X be the universal cover of Y. So, X is simply connected and F freely acts on X. This action induces an action of G on X. Since F acts on X freely, the isotropy groups of vertices and edges under the action of G are all isomorphic to the group H, which is finitely presented. Since Y is finite, the number of orbits of cells (under the action of either F or G) is finite. Hence we may apply Theorem 4.3.A and conclude that G is finitely presented. □

LEMMA 4.3.C. *Let H be a subgroup of finite index of a group G. Then H is finitely presented if and only if G is finitely presented.*

PROOF. Suppose first that G is finitely presented. Let Y be a finite CW-complex with $\pi_1(Y) = G$, and let X be the universal cover of Y. Then H acts freely on X and the number of the orbits of the cells is finite because $Y = X/G$ is finite and the index $[G : H]$ is finite. Theorem 4.3.A now implies that H is finitely presented.

[2] in [129, p. 37], on the second line $(a_0, \ldots, a_i, a_{i+1}, \ldots, a_n)$ should be replaced by $(a_0, \ldots, a_i, a_{i+2}, \ldots, a_n)$.

Suppose now that H is finitely presented. Let

$$H' = \bigcap_{g \in G} gHg^{-1}.$$

This intersection is actually finite, because $[G : H] < \infty$, and hence H' is of finite index in both G and H. By the already proved part of the lemma, H' is finitely presented. Since H' is obviously normal in G, we have an exact sequence of groups $1 \to H' \to G \to G/H' \to 1$. Since G/H' is finite, it is finitely presented. Now Lemma 4.3.B implies that G is finitely presented. □

THEOREM 4.3.D. Mod_S *is finitely presented for any compact* (*orientable*) *surface* S.

PROOF. In view of Lemma 4.3.C it is sufficient to prove that the groups PMod_S are finitely presented. In order to deal with PMod_S we will use double induction on the genus and the number of boundary components, as in the proof of Theorem 4.2.C.

Suppose that $\partial S \neq \emptyset$. Let R be the result of gluing a disc to one of the boundary components of S. By the inductive assumption PMod_R is finitely presented. The kernel of the natural homomorphism $\rho : \mathrm{PMod}_S \to \mathrm{PMod}_R$ from 2.8 is finitely presented (by Theorem 2.8.C this kernel is isomorphic to $\pi_1(R)$ with few exceptions; these exceptions are easy to deal with directly), and hence Lemma 4.3.C implies that PMod_S is also finitely presented. This completes the induction step of the induction on the number of boundary components.

Next, suppose that S is closed. If S is a sphere, then $\mathrm{PMod}_S = 1$. If S is a torus, then PMod_S is isomorphic to $\mathrm{SL}_2(\mathbf{Z})$. It is well known that the latter group is finitely presented (cf. [214], for example). So, we may assume that the genus of S is $\geqslant 2$. In this case we are going to apply Theorem 4.3.A to the action of PMod_S on the geometric realization of $C(S)$. We need to verify the assumptions of Theorem 4.3.A. Theorem 3.2.C implies (i) (cf. also the discussion preceding Theorem 3.2.C). Let us now check (iii). Suppose that $\langle C_1 \rangle$, $\langle C_2 \rangle$ are connected by an edge of $C(S)$, and let G be the isotropy group of (the geometric realization of) this edge. We may assume that $C_1 \cap C_2 = \emptyset$. Let G_0 be the subgroup of PMod_S consisting of the isotopy classes of diffeomorphisms $F : S \to S$, such that $F(C_1) = C_1$, $F(C_2) = C_2$ and F preserves the sides of both circles C_1, C_2. Clearly, G_0 is a subgroup of finite index in G (in fact, the index is $\leqslant 4$). Such diffeomorphisms F can be cut along $C_1 \cup C_2$, and this leads to a canonical homomorphism $G_0 \to \mathrm{PMod}_Q$, where Q is the result of cutting S along $C_1 \cup C_2$. Clearly, this homomorphism is surjective, its kernel is generated by the Dehn twists about C_1 and C_2 (the latter assertion follows from Corollary 4.1.B applied twice), and these Dehn twists commute (by Lemma 4.1.F(i)). Since S is closed, the genus of the components of Q is (strictly) less than the genus of S. So, the inductive assumption implies that PMod_Q is finitely presented. By Lemma 4.3.B, G_0 is also finitely presented. Finally, by Lemma 4.3.C the isotropy group G is finitely presented. So, we proved an assertion even stronger than (iii). The proof of (ii) is similar and simpler. In order to check (iv), notice that the number of orbits of cells of dimension $\leqslant 2$ is nothing more than the number of arrangements of $\leqslant 3$ disjoint circles on S considered up to a diffeomorphism. Clearly, this number is finite. So, we can apply Theorem 4.3.A and conclude that PMod_S is finitely presented. This completes the induction step of the induction on genus, and hence the proof. □

Theorem 4.3.D was originally proved by McCool [161] by methods of the combinatorial group theory. His approach allows, in principle, computation of an explicit finite presentation of Mod_S for a given S, but this turned out to be too difficult to be done for any S. Later on, Hatcher and Thurston [91] suggested a geometric approach to the finite presentability and to a construction of an explicit finite presentation. The above proof was suggested by the author in [98]. It has some features in common with [91], such as the use of families of functions in order to prove the simply connectivity of an appropriate complex (cf. the proof of Theorems 3.2.B and 3.2.C above), and the use of Theorem 4.3.A (only implicit in [91]). In some sense, this proof can be considered as an ultimate simplification of the proof of Hatcher and Thurston (at least for the time being). One more approach to this theorem is indicated in 5.4.

The complex used in [91] is very complicated, and Hatcher and Thurston did not achieve their goal of writing down an explicit finite presentation (that does not diminishes the significance of their exceptionally beautiful paper!). Their complex was somewhat simplified by Harer [75], and using this simplified complex Wajnryb [230] succeeded in writing down an explicit finite presentation for closed surfaces and for surfaces with one boundary component. The proof involves some extremely complicated computations. The reader should be warned that [230] contains some mistakes, even in the statements of the main theorems, corrected much later in [21]. Recently, Wajnryb [232] provided a new exposition of his work. In addition, [232] contains a new proof of the simply connectivity of the Hatcher–Thurston complex.

Let us briefly describe the Wajnryb presentation for closed surfaces. As generators he uses the Humphries generators of Corollary 4.2.F. Relations are the following. First, for any pair of disjoint circles from the list of Corollary 4.2.F, the Dehn twists about them are related by the commutation relation of Lemma 4.1.F(i). Second, for any pair of circles from this list intersecting (transversely) at one point, the Dehn twists about them are related by the Artin relation of Lemma 4.1.F(ii). There are three additional relations. In order to describe them, let us, as usual, denote the isotopy classes of the circles $A_1, \dots, A_g, B_1, \dots, B_{g-1}, C_1, \dots, C_g$ in Figure 7 by $\alpha_1, \dots, \alpha_g, \beta_1, \dots, \beta_{g-1}, \gamma_1, \dots, \gamma_g$, respectively.

In order to describe the first relation, we need a circle D_2 on S intersecting transversely A_2 at exactly one point, not intersecting any other circle in Figure 7, and not isotopic to C_2 (up to isotopy D_2 is characterized by these properties; it is situated similarly to C_2, but is contained in the lower part of S). Let $\delta_2 = \langle D_2 \rangle$. Then $\delta_2 = u(\gamma_2)$, where $u = t_{\alpha_2} t_{\beta_1} t_{\alpha_1} t_{\gamma_1}^2 t_{\alpha_1} t_{\beta_1} t_{\alpha_2}$, as one may check straightforwardly. Hence, $t_{\delta_2} = u t_{\gamma_2} u^{-1}$. It turns out that $(t_{\gamma_1} t_{\alpha_1} t_{\beta_1})^4 = t_{\gamma_2} t_{\delta_2}$, and using our formulas for t_{δ_2} and u, we can express this relation in terms of the Humphries generators. This is the first additional relation.

The next relation is, essentially, the lantern relation of Lemma 4.1.H, expressed in terms of the Humphries generators. To write it down, Wajnryb chose an embedding of S_0^4 in S such that the boundary circles C_0, C_1, C_2, C_3 of S_0^4 are mapped to the circles C_1, B_2, B_3, C_3 of Figure 7. Then he expressed the Dehn twists about the images of the

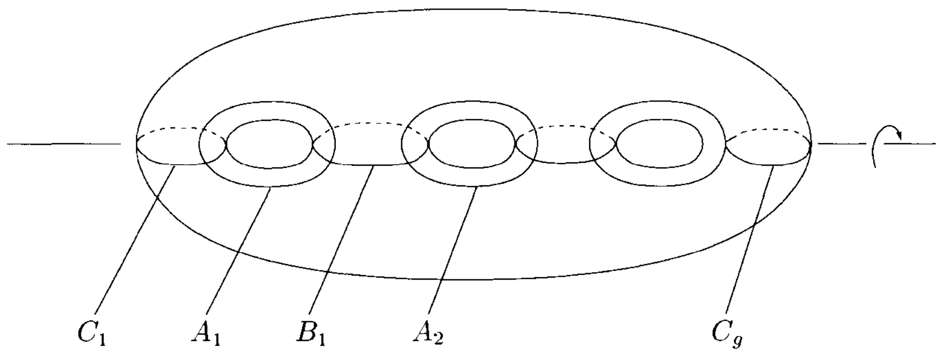


Fig. 12.

circles C_{12}, C_{13}, C_{23}, and the Dehn twist about C_3 in terms of the Humphries generators. After this he wrote the lantern relation in terms of these expressions. This is the second additional relation.

The last relation is concerned with the so-called *hyperelliptic involution* of S. It is an orientation-preserving diffeomorphism $H : S \to S$ of order 2 (i.e., $H \circ H = \mathrm{id}_S$). Up to isotopy, it is characterized by the following properties: it maps each of the $2g + 1$ circles $C_1, A_1, B_1, A_2, B_2, \ldots, B_{g-1}, A_g, C_g$ to itself, and reverses the orientation of all these circles (and hence interchanges the components of the complement in S of the union of these $2g + 1$ circles). It can be represented by a 180° rotation around the horizontal axis of symmetry in Figure 12. The isotopy class h of H is also called the *hyperelliptic involution*. (Note that, as it is described here, the hyperelliptic involution depends on the choice of a collection of circles diffeomorphic to the one in Figure 7, and different choices lead to different, but conjugate, hyperelliptic involutions. For the purposes of writing down the Wajnryb presentation, we need only one such collection of circles.) Since $H(C_g) = C_g$ and H is orientation-preserving, $ht_{\gamma_g}h^{-1} = t_{\gamma_g}$ by Lemma 4.1.C, or, what is the same, $ht_{\gamma_g} = t_{\gamma_g}h$. The third additional relation is, essentially, this commutation relation. It turns out that $h = t_{\gamma_g} f t_{\gamma_g}$, where $f = t_{\alpha_g} t_{\beta_{g-1}} \ldots t_{\beta_1} t_{\alpha_1} t_{\gamma_1}^2 t_{\alpha_1} t_{\beta_1} \ldots t_{\beta_{g-1}} t_{\alpha_g}$. Clearly, h commutes with t_{γ_g} if and only if f does. Wajnryb expressed t_{γ_g} in terms of the Humphries generators, and then, using this expression and the above formula for f, he wrote the commutation relation $ft_{\gamma_g} = t_{\gamma_g} f$ in terms of the Humphries generators. This is his third (and the last) additional relation.

The Wajnryb presentation for surfaces with one boundary component is obtained from the Wajnryb presentation for closed surfaces by omitting the third additional relation. (If we made a hole in the surface on Figure 7 at the right end of it, the Dehn twist about the same circles serve as generators.)

The above description of the Wajnryb presentations in geometrical terms is, essentially, borrowed from Birman's survey [18]. (Note that the relation (6) in [18] is incorrect as stated. One should replace in it the element G by the above expression for the hyperelliptic involution h.)

Now, let S be a closed surface of genus 2. In this case the second additional relation follows from the other relations. The first finite presentation of Mod_S in this case was discovered by Bergau and Mennicke [12]. Their proof turned out to be incomplete, and their presentation was justified only about ten years later by Birman and Hilden [19]. Instead of the first two additional relations of the Wajnryb presentation, this presentation contains the following two relations:

$$(t_{\gamma_1} t_{\alpha_1} t_{\beta_1} t_{\alpha_2} t_{\gamma_2})^6 = 1;$$
$$\left(t_{\gamma_2} t_{\alpha_2} t_{\beta_1} t_{\alpha_1} t_{\gamma_1}^2 t_{\alpha_1} t_{\beta_1} t_{\alpha_2} t_{\gamma_2}\right)^2 = 1$$

(the second one expresses the fact that the hyperelliptic involution has order 2). A special feature of this case is that the hyperelliptic involution h commutes with all 5 generators of Theorem 4.2.D or Corollary 4.2.F (known already to Dehn [45]), as is clear from the above description of h, and hence belongs to the center of Mod_S. In fact, it generates the center in this case. On the contrary, if the genus is $\geqslant 3$, then h does not commutes with t_{γ_2}, and, in fact, the center is trivial. Cf. 7.5.

For surfaces S of genus 0 (i.e., a spheres with holes), Mod_S is closely related to the classical Artin braid groups. This allows to write down a presentation of Mod_S in this case. For the details, see [16, Chapter 4].

Recently, Gervais [60] used the Wajnryb presentation and induction on the number of the boundary components in order to derive an explicit (and very symmetric) presentation of Mod_S for surfaces S with an arbitrary number of boundary components.

Finally, let us mention a related work of Gervais [59], where he derived some *infinite* presentations of mapping class groups, similar to the Wajnryb presentation, but having as generators all Dehn twists or all Dehn twists about nonseparating circles. His presentation with all Dehn twists as generators was recently simplified by Luo [146].

5. Teichmüller spaces

Teichmüller spaces provide another class of fundamental geometric objects on which mapping class groups act. Our approach to Teichmüller spaces is closer in spirit to that of Fricke [55] than to the approach of Teichmüller himself. In particular, we define the Teichmüller spaces in terms of metrics of constant curvature, and not in terms of Riemann surfaces as Teichmüller did. In fact, the work of Fricke precedes the work of Teichmüller by about forty years. Probably, Fricke was also the first to define the mapping class groups.

In our discussion we stress the connection of Teichmüller spaces with complexes of curves (and especially the results of Harvey [86,87]; cf. 5.4) and other combinatorial objects, namely, the so-called ideal triangulations (cf. 5.5). As the first application, we will indicate another approach to the finite presentability of the mapping class groups (cf. Corollaiy 5.4.B). In the next section we will present a much deeper application of Teichmüller spaces to the mapping class groups, namely, the computation of the virtual cohomological dimension of the mapping class groups (cf. Theorems 6.4.A, 6.4.B and 6.4.C).

5.1. *Definitions*

Let us assume that our (compact orientable) surface S admits a Riemannian metric of constant curvature -1 with geodesic boundary. As is well known (see 2.5), such metrics exist if and only if S has negative Euler characteristic. Let us fix some number $l > 0$ and consider only metrics such that all components of the boundary ∂S have length l. The choice of l is largely irrelevant, but at a later point it will be convenient to have l small. Let us denote by H_S the set of such metrics. We can endow H_S with some convenient topology, say the C^∞-topology, by considering Riemannian metrics as tensor fields. Since diffeomorphisms act on tensor fields (by the push-forward), the group Diff(S) acts on H_S. The quotient space $H_S/\operatorname{Diff}_0(S)$ by the subgroup of diffeomorphisms isotopic to the identity is called the *Teichmüller space* of S and is denoted by T_S. Clearly, the quotient group $\operatorname{Mod}_S = \operatorname{Diff}(S)/\operatorname{Diff}_0(S)$ acts on T_S and the quotient space T_S/Mod_S is equal to $H_S/\operatorname{Diff}(S)$, which is called the *moduli space of hyperbolic structures* on S and is denoted by M_S. Hence $T_S/\operatorname{Mod}_S = M_S$.

Sometimes (cf., for example, 5.5) it is more convenient to consider another version of Teichmüller spaces, namely the one related to so-called punctured surfaces. By a *punctured surface* R we understand a closed (orientable) surface *together* with a finite subset of it. The points of this finite subset are called *punctures* and we denote by R° the result of deleting all punctures from the surface. Let R be a punctured surface such that R° admits a complete Riemannian metric of constant curvature -1. Again, it is well known that such metrics exist if and only if R° has negative Euler characteristic. We denote by H_R the set of such metrics. This set also can be endowed with a natural topology. This time, the group $\operatorname{Diff}(R^\circ)$ acts on H_R, and the quotient space $H_S/\operatorname{Diff}_0(R^\circ)$ by the subgroup of diffeomorphisms isotopic to the identity is called the *Teichmüller space* of R and is denoted by T_R. Naturally, the group $\operatorname{Mod}_R = \operatorname{Diff}(R^\circ)/\operatorname{Diff}_0(R^\circ)$ is called the *mapping class group* of R. It naturally acts on T_R and the quotient space T_R/Mod_S is equal to $H_R/\operatorname{Diff}(R^\circ)$, which is called the *moduli space of hyperbolic structures* on R and is denoted by M_R. Clearly, $T_R/\operatorname{Mod}_R = M_R$.

Note that there is no real difference between mapping class groups of compact and punctured surfaces. If S is a compact surface, then we can collapse each component of ∂S into a point (one for each boundary component) and get a closed surface. Images of boundary components form a finite subset of this surface, and the resulting surface together with this finite subset is a punctured surface. Let us denote it by R. Clearly, $S \setminus \partial S = R^\circ$. This leads to a natural map $\operatorname{Diff}(S) \to \operatorname{Diff}(R^\circ)$. The homomorphism $\operatorname{Mod}_S \to \operatorname{Mod}_R$ induced by it is easily seen to be an isomorphism.

5.2. *Length functions and Fenchel–Nielsen flows*

Let α be a vertex of $C(S)$, i.e., the isotopy class of a nontrivial circle. For any metric $h \in H_S$ the isotopy class α contains exactly one geodesic with respect to the metric h. Let us denote by $l_\alpha(h)$ the length of this geodesic. Clearly, this length depends only on the image of h in T_S. Hence, we can consider l_α as a function $T_S \to \mathbf{R}_{>0}$. We call this function the *length function* corresponding to α. For a simplex $\sigma = \{\alpha_1, \dots, \alpha_m\}$ of $C(S)$ we may consider the map $L_\sigma : T_S \to \mathbf{R}^m_{>0}$ given by the formula $L_\sigma(x) = \{l_{\alpha_1}(x), \dots, l_{\alpha_m}(x)\}$. Clearly, it comes from a similar map $L_\sigma : H_S \to \mathbf{R}^m_{>0}$. The map L_σ is especially important when σ is a simplex of maximal dimension. In this case it turns out to be a principal bundle over $\mathbf{R}^m_{>0}$ having as the structure group the additive group $\mathbf{R}^m$.

Let us describe the corresponding action of $\mathbf{R}^m$ on T_S. Let $h \in H_S$ be a metric representing a point $x \in T_S$, and let C_i be the geodesic circle with respect to h in the isotopy class α_i ($1 \leqslant i \leqslant m$). If we cut S along this geodesic C_i, we get a new surface (with a metric of curvature -1 and geodesic boundary), which has two new boundary components resulting from C_i. In order to restore S from this new surface, we have to glue back these two boundary components. Clearly, these boundary components should be glued by an isometry, but this does not define the gluing uniquely. Two isometric gluings may differ by a rotation. It seems at first sight that the set of all possible isometric gluings is parameterized by a circle. In fact, since we are interested in the resulting point of T_S (and not only of M_S), i.e., we are interested in the resulting metric up to isotopy (i.e., up to the action of $\mathrm{Diff}_0(S)$) and not only up to isometry, the set of all possible gluings is parameterized by the universal cover of this circle, i.e., by a copy of $\mathbf{R}$. More precisely, there are (at least) two natural identifications of this universal cover with $\mathbf{R}$: the one corresponding to the measuring of the amount of rotation by the angle, and the other corresponding to the measuring of the amount of rotation by the length (one can get from the first to the second by multiplication by $l_{\alpha_i}(h)/2\pi$). Another ambiguity comes from the need to orient the circle C_i in order to get the identification, but none of these is of any importance for us now. To summarize, we may cut the original surface and then glue it back by a gluing differing from the one leading back to the original surface by a parameter $r \in \mathbf{R}$. This defines a map $\mathbf{R} \times T_S \to T_S$, and it is easy to see that this map is an action of the additive group $\mathbf{R}$. This action is usually called a *Fenchel–Nielsen flow*. Of course, this action depends on the vertex α_i. For any $h \in H_S$, the corresponding circles C_i, $1 \leqslant i \leqslant m$, are disjoint (because σ is a simplex), and this easily implies that the actions corresponding to different vertices of σ commute. Hence, they define an action of $\mathbf{R}^m$.

The fact that this action turns L_σ into a principal bundle amounts, essentially, to the following. In the notation of the previous paragraph, if the simplex σ is maximal and we cut S along all circles C_i, $1 \leqslant i \leqslant m$, we will get a collection of discs with two holes. A metric of the constant curvature -1 with geodesic boundary on a disc with two holes is uniquely determined, up to isotopy, by the lengths of its boundary components. Hence, the cut surface is determined by $L_\sigma(x)$. Clearly, if two surfaces (with metrics) lead to the same cut surface, then they differ only by the gluing, i.e., by the action of $\mathbf{R}^m$. It follows that the orbits of $\mathbf{R}^m$ are exactly the fibers of the map L_σ. One needs also to check that the action of $\mathbf{R}^m$ on T_S is free, and also the existence of local sections. We refer to [53, Exposé 7] for this.

5.3. *The topology of Teichmüller spaces*

As we just saw, the Teichmüller space T_S is the total space of a principal bundle over $\mathbf{R}^m_{>0}$ with the fiber $\mathbf{R}^m$. Since $\mathbf{R}^m_{>0}$ is contractible, it follows that T_S is homeomorphic to $\mathbf{R}^m_{>0} \times \mathbf{R}^m$, and hence to $\mathbf{R}^{2m}$. The number m is the maximal number of pairwise disjoint pairwise non-isotopic non-trivial circles on S, hence $m = 3g - 3 + b$, where g is the genus and b is the number of boundary components of S. We conclude that T_S is homeomorphic to $\mathbf{R}^{6g-6+2b}$.

In addition, one can introduce a smooth structure on T_S in a number of ways, all leading to the same smooth structure. This smooth structure can be characterized by the fact that all length functions L_α are smooth. In fact, a finite number of them is sufficient to embed T_S in a Euclidean space $\mathbf{R}^n$ as a smooth submanifold. The straightforward approach requires $9g-9+3b$ length functions (cf. [53, Exposés 7 and 11]). But for a closed S it is sufficient to use only $\dim T_S+1$ length functions. And if we consider another version of the Teichmüller space T_S, in which the lengths of the boundary components are not fixed as in 5.1, then for *non-closed* S it is sufficient to use only $\dim T_S$ length functions. See Schmutz [209] and the references there for these results.

No matter what definition of the smooth structure is used, it is easy to see that the principal bundle from 5.2 is smooth. It follows that T_S is not only homeomorphic, but also diffeomorphic to a Euclidean space (no results about topological manifolds need to be used!). Also, the Fenchel–Nielsen flows are smooth. Moreover, we can replace *smooth* by *real analytic* everywhere. For closed S the Teichmüller space T_S even has a canonical complex analytic structure, but it plays no role in the sort of questions we are interested in (applications to the mapping class groups).

The discussion in 5.2 and in this section applies equally well to Teichmüller spaces of punctured surfaces. In particular, Teichmüller space T_R of a punctured surface R is diffeomorphic to $\mathbf{R}^{6g-6+2b}$, where g is the genus of R and b is the number of punctures, unless R is a torus without punctures, in which case T_R is diffeomorphic to $\mathbf{R}^2$.

5.4. *Corners and the mapping class groups*

As we noticed in 5.1, Mod_S acts on T_S. This action is not free, because Mod_S has torsion and T_S is homeomorphic to a Euclidean space (any element of finite order has a fixed point in T_S, say by the Smith theory – see, for example, [53, Exposé 11, end of §IV]). But it is as close to being free as possible: any torsion free subgroup Γ of Mod_S acts freely on T_S (the argument goes as follows: if the isotopy class of a diffeomorphism F fixes a point of T_S represented by a metric $h \in H_S$, then F is isotopic to an isometry F' of the metric h; it is well known that the group of isometries of a metric of constant curvature -1 on S is finite). It is well known that Mod_S contains torsion free subgroups, even of finite index. In fact, the kernel of the natural homomorphism $\mathrm{Mod}_S \to \mathrm{Aut}(H_1(S, \mathbf{Z}/m\mathbf{Z}))$ defined by the action of diffeomorphisms on homology is such a subgroup for any $m \geqslant 3$. This result is a combination of a theorem of Serre (cf. Theorem 6.8.A) and a classical result of Nielsen (cf. Theorem 7.1.A) about realization of finite cyclic subgroups of Mod_S by subgroups of $\mathrm{Diff}(S)$; see Corollary 7.1.B. In addition, this action is properly discontinuous. Hence, for torsion free Γ the quotient space T_S/Γ is a $K(\Gamma, 1)$-space and, of course, a manifold. It turns out that this manifold is never compact, but one can construct a compact manifold from it if Γ is of finite index in Mod_S either by adding some boundary at infinity or by deleting a neighborhood of infinity. Moreover, one can do this in a fairly canonical way.

Before we describe these constructions, we note that the resulting compact manifolds with boundary will *not* be smooth manifolds in the usual sense, but smooth manifolds with corners. Recall that a *smooth manifold with corners* is simply a smooth manifold modeled on the products $\mathbf{R}^n_{\geqslant 0} \times \mathbf{R}^m$ for all n, m (and not only for $n = 1$ as manifolds with boundary). The simplest type of such manifolds is presented by the subsets of usual smooth manifolds locally described by inequalities of the form $x_1 \geqslant 0, \dots, x_n \geqslant 0$, where $(x_1, \dots, x_n, \dots, x_{n+m})$ is a chart.

Let us fix some small $\varepsilon > 0$ and let us assume (only for convenience) that the boundary length l fixed in 5.1 is equal to ε. Let $T_S(\varepsilon)$ be the subset of T_S defined by the inequalities $l_\alpha(x) \geqslant \varepsilon$ for all isotopy classes α of non-trivial circles. Clearly, $T_S(\varepsilon)$ is invariant under the action of Mod_S on T_S. The basic properties of $T_S(\varepsilon)$ are summarized in the following theorem.

THEOREM 5.4.A. *If ε is sufficiently small, then the following holds.*

(i) *The action of* Mod_S *on $T_S(\varepsilon)$ is properly discontinuous.*

(ii) *Any torsion free subgroup Γ of* Mod_S *acts on $T_S(\varepsilon)$ freely, and its quotient space $T_S(\varepsilon)/\Gamma$ is a smooth manifold with corners.*

(iii) *The quotient space $T_S(\varepsilon)/\mathrm{Mod}_S$ is compact.*

(iv) *$T_S(\varepsilon)$ is contractible.*

(v) *The boundary $\partial T_S(\varepsilon)$ is homotopy equivalent to the (geometric realization of the) complex of curves $C(S)$.*

PROOF. First, the ε should be chosen so small that the following holds: if h is a metric of constant curvature -1 on S and C_1, C_2 are two geodesic circles (with respect to h) on S of length $\leqslant \varepsilon$, then C_1, C_2 are either equal or disjoint. The existence of such an ε is well known; cf. [1, Chapter II], corollary of Lemma 1 in Section 3.3; it also follows from [29, Corollary 4.1.2]. The property (i) and the first part of the property (ii) immediately follow from the corresponding properties of T_S itself. Property (iii) follows from the so-called Mumford compactness theorem [187]; see [29, Theorem 6.6.5] or [1, Chapter II, theorem in Section 3.4] for a closely related result. The proof of (iv) is the most technical part of the proof of the theorem; we refer to [106, Theorem 3] for a proof. Next, we turn to the property (v). For any vertex α of $C(S)$, let us consider the subset $T_S(\varepsilon, \alpha) = \{x \in T_S(\varepsilon) \mid l_\alpha(x) = \varepsilon\}$ of $T_S(\varepsilon)$. This subset turns out to be closely related to $T_{S_\alpha}(\varepsilon)$, where S_α is the result of cutting S along a representative of α, and one can use this together with (iv) for S_α in order to prove that $T_S(\varepsilon, \alpha)$ is contractible. It is easy to see that the boundary of $T_S(\varepsilon)$ is equal to the union of subsets $T_S(\varepsilon, \alpha)$. By the choice of ε, any collection $T_S(\varepsilon, \alpha_i)$ of such subsets has a nonempty intersection if and only if the vertices α_i can be represented by disjoint circles. Hence, $C(S)$ is the nerve of a cover of $\partial T_S(\varepsilon)$ by contractible subsets. The property (v) follows. Finally, the second part of the property (ii) follows from the fact that $T_S(\varepsilon)$ itself is a manifold with corners. This follows, in turn, from the fact that in a neighborhood of a point $x \in \partial T_S(\varepsilon)$ the subset $T_S(\varepsilon)$ is defined by a *finite* number of the inequalities of the form $l_{\alpha_i} \geqslant \varepsilon$ and, moreover, the corresponding α_i form a simplex, and hence can be included as coordinate functions in a chart (by 5.2).

A more detailed outline of the proof is contained in [98, Section 4], with complete details provided in [106]. An outline of another proof is contained in [79, Section 3]. □

COROLLARY 5.4.B. Mod_S *is finitely presented.*

PROOF. Let Γ be a torsion free subgroup of finite index in Mod_S. It follows from (i) and (iv) that $T_S(\varepsilon)/\Gamma$ is a $K(\Gamma, 1)$-space. Because Γ is of finite index in Mod_S, it follows from (iii) that $T_S(\varepsilon)/\Gamma$ is compact. In view of (ii), $T_S(\varepsilon)/\Gamma$ is a smooth manifold with corners. The last property implies that $T_S(\varepsilon)/\Gamma$ is homotopy equivalent to a finite

CW-complex (moreover, it admits a finite triangulation). Hence, there exists a finite CW-complex which is a $K(\Gamma, 1)$-space. It follows that Γ is finitely presented. By Lemma 4.3.C, Mod_S is also finitely presented. □

The idea behind Theorem 5.4.A is due to Harvey [86,87]. In [87], Harvey announced a construction of a manifold with corners X_S, containing the Teichmüller space T_S as its interior and such that the properties (i)–(v) of Theorem 5.4.A hold with $T_S(\varepsilon)$ replaced by X_S. He was motivated by the analogy (first noticed also by him) between the mapping class groups and arithmetic groups (cf. 9.1) and by the work of Borel and Serre [23]. Adding the corners to T_S in this manner turned out to be a more delicate matter than excising a part from T_S as in Theorem 5.4.A. A proof of the existence of such an X_S was provided in [105], together with a geometric interpretation of the points in the complement $X_S \setminus T_S = \partial X_S$.

5.5. *Ideal triangulations*

Let R be a punctured surface in the sense of 5.1 with a *non-empty* set of punctures. An *ideal triangulation* of R is a collection of (embedded) arcs with disjoint interiors connecting the punctures of R such that if we cut R along all these arcs, we will get a collection of triangles (with sides resulting from these arcs). The word *ideal* is justified by the standard understanding that the punctures are points at the infinity. In other words, we ideally triangulate R° rather than R. Note also that two endpoints of an arc are permitted to coincide and a triangle of an ideal triangulation is *not* determined by the set of its vertices (which are punctures) in general.

The ideal triangulations, considered up to isotopy, are exactly the top dimensional simplices of a simplicial complex $A(R)$, which we now define. The vertices of $A(R)$ are the isotopy classes of non-trivial arcs connecting punctures of R. Here an arc is called *trivial* if two its endpoints coincide and the resulting circle bounds a disc in R which does not contain other punctures. The isotopies are permitted, naturally, only in the class of arcs connecting punctures, so they are fixed on the endpoints. The simplices are defined in a manner similar to the simplices of complexes of curves: a collection of vertices forms a simplex if these vertices can be represented by arcs with disjoint interiors. Like complexes of curves, $A(R)$ is a *flag complex* (cf. 3.1). As in the case of complexes of curves, this can be established by using complete metrics of constant curvature -1, this time on R°, and geodesic arcs. Clearly, Mod_R acts on $A(R)$.

We also need a subcomplex $A_\infty(R)$ of $A(R)$. It has the same vertices as $A(R)$. A collection of vertices forms a simplex of $A_\infty(R)$ if those vertices can be represented by arcs with disjoint interiors such that the result of cutting R along these arcs has at least one component which is not a disc meeting the set of punctures only in its boundary. In other words, simplices of $A(R)$ which are *not* simplices of $A_\infty(R)$ correspond to the isotopy classes of *ideal cell decompositions* of R in an obvious sense.

THEOREM 5.5.A. *Let Δ^{b-1} be the* open *standard $(b-1)$-dimensional simplex, where b is the number of punctures of R. Let us endow Δ^{b-1} with the trivial action of* Mod_R, *and the* *product* $T_R \times \Delta^{b-1}$ *with the product action. Then* $T_R \times \Delta^{b-1}$ *is* Mod_R*-equivariantly homeomorphic to the geometric realization of* $A(R)$ *with the geometric realization of* $A_\infty(R)$ *removed.*

Note that the homeomorphism implied in this theorem is *not* canonical: different proofs lead to apparently different homeomorphisms.

The early history of this theorem is not well documented in the literature. According to [79, Chapter 2], the idea of this triangulation is due to Thurston; his approach was based on hyperbolic geometry. The first published account is contained in [78], where an alternative proof, due to Mumford and based on work of Strebel on quadratic differentials on Riemann surfaces (cf. [218,219]), was outlined. Proofs based on hyperbolic geometry were published by Penner [197] and by Bowditch and Epstein [24]. Penner's approach seems to be more remote from the original ideas of Thurston in that the Lorentz model of the hyperbolic plane plays a crucial role in it. His ideas were extended, to some degree, to the higher dimensions [50] and to closed surfaces [189], and in [202] Penner applied them to the so-called *universal Teichmüller space*. Unfortunately, all this falls outside the scope of the present paper.

The complexes $A(R)$ and $A_\infty(R)$ are, together with complexes of curves, the basic combinatorial objects of the topology of surfaces. To a large extent they are simpler than complexes of curves, Theorem 5.5.A being the main reason, but they are not defined for *closed* surfaces. Among the main applications of the complexes $A(R)$ and Theorem 5.5.A are the computation of the virtual Euler characteristic of the mapping class group by (independently) Harer and Zagier [85] and Penner [198,199] (cf. 6.7), computations and estimates of the Weil–Petersson volumes of Teichmüller spaces by Penner [201], Theorem 8.1.A about automorphisms of complexes of curves (cf. 8.4), the proof of Witten's conjecture [233] by Kontsevich [126] and the theorem of Mosher [181,182] to the effect that the mapping class groups are *automatic* in the sense of [49]. For an exposition of the work of Kontsevich we refer to [141,142], and for an introduction to Mosher's results we refer to his own survey [183].

As an easy and useful corollary of Theorem 5.5.A we point out the following result.

COROLLARY 5.5.B. *Every simplex of* $A(R)$ *is a face of a top dimensional simplex. Every codimension* 1 *simplex of* $A(R)$ *is a face of* 1 *or* 2 *top dimensional simplices. Any two top dimensional simplices* Δ, Δ' *of* $A(R)$ *can be connected by a chain of simplices* $\Delta = \Delta_1, \dots, \Delta_m = \Delta'$ *such that any two consecutive simplices* Δ_i, Δ_{i+1} *have a common codimension* 1 *face.*

A much more elementary proof of this corollary was provided by Hatcher [89]; cf. [89, corollary].

6. Cohomological properties

This section is devoted to cohomological properties of mapping class groups. It is centered around the computation of the so-called virtual cohomological dimension of mapping class groups (see 6.4). As a preparation for this computation, we present in 6.1 a very short and idiosyncratic introduction to the cohomology of groups. This introduction is heavily slanted toward phenomena crucial for understanding of both mapping class groups and arithmetic groups (we will discuss the latter in 9.1) and known as the Bieri–Eckmann duality (cf. [13]). The latter is actually a manifestation of the Poincaré–Lefschetz duality, as we try to explain in 6.1. After this introduction, we prove in 6.2–6.4 Harer's theorem [76,78] computing the virtual cohomological dimension of the mapping class groups. In 6.5–6.7 we review, mostly without proofs, some related cohomological properties of mapping class groups. Finally, in 6.8 we review some results about cohomology of the mapping class groups with finite coefficients.

6.1. *Cohomology of groups and the Poincaré–Lefschetz duality*

The goal of this section is to present an introduction to cohomology of groups and a fragment of the theory of Bieri and Eckmann [13] (see, especially, Theorem 6.1.H).

Let Γ be a discrete group and let X be a $K(\Gamma, 1)$-space with a fixed isomorphism $\Gamma \to \pi_1(X)$, which will be used to identify these groups. The homology or cohomology of the group Γ can be defined as the homology or cohomology of the space X. In order to capture the essential properties of Γ, one has to consider also homology and cohomology with so-called *twisted* coefficients. Let M be a Γ-module (in other words, M is a module over the group ring $\mathbf{Z}[\Gamma]$). Then the homology $H_*(\Gamma, M)$ and cohomology $H^*(\Gamma, M)$ of Γ with (twisted) coefficients M can be defined as follows.

For a space Y let us denote by $C_*(Y)$ its singular complex, or, if Y is a CW-complex, its complex of cellular chains. Let $X^\sim \to X$ be the universal covering of X. The fundamental group $\Gamma = \pi_1(X)$ acts naturally on $X^\sim$, turning $C_*(X^\sim)$ into a complex of Γ-modules. By definition,

$$H_*(\Gamma, M) = H_*(X, M) = H^*\big(C_*(X^\sim) \otimes_\Gamma M\big),$$

where $\otimes_\Gamma$ means the tensor product over the group ring $\mathbf{Z}[\Gamma]$. If $M = \mathbf{Z}$ with the trivial structure of Γ-module (i.e., $\gamma m = m$ for all $\gamma \in \Gamma$, $m \in M$), then

$$C_*(X^\sim) \otimes_\Gamma M = C_*(X^\sim) \otimes_\Gamma \mathbf{Z} = C_*(X)$$

and we recover the usual homology of Γ and X. (In fact, if two singular simplices or cells c_1, c_2 in $X^\sim$ are projected to the same simplex or cell in X, then $c_1 = \gamma c_2$ for some $\gamma \in \Gamma$, and hence $c_1 \otimes m = \gamma c_2 \otimes m = c_2 \otimes \gamma m = c_2 \otimes m$ for any $m \in M$.) Cohomology is defined in a similar way:

$$H^*(\Gamma, M) = H^*(X, M) = H^*\big(\mathrm{Hom}_\Gamma\big(C_*(X^\sim), M\big)\big),$$

where $\mathrm{Hom}_\Gamma(C_*(X^\sim), M) \subset C^*(X^\sim, M)$ is the group of *equivariant* M-valued cochains, i.e., the group of cochains f such that $f(\gamma c) = \gamma f(c)$ for any $\gamma \in \Gamma$ and any chain c. As above, if $M = \mathbf{Z}$ with the trivial structure of Γ-module, we recover the usual cohomology of Γ and X.

If we choose a lift to $X^{\sim}$ of any singular simplex or cell in X, then any equivariant cochain will be defined by its values on these lifts. It follows that $\mathrm{Hom}_{\Gamma}(C_*(X^{\sim}), M)$ and $C^*(X, M)$ are isomorphic as graded Abelian groups. But the differentials are different (in the first complex, the differential is *twisted*).

Now, let us define the *cohomological dimension* $\mathrm{cd}\,\Gamma$ of Γ as the supremum (which can be ∞) of the numbers n such that $H^n(\Gamma, M) \neq 0$ for some Γ-module M.

THEOREM 6.1.A. *If X is a closed topological manifold of dimension n, then* $\mathrm{cd}\,\Gamma = n$.

PROOF. First, note that $H^m(X, M) = 0$ for $m > n$ by same reasons as for trivial coefficients. Thus we only need to find a Γ-module M such that $H^n(X, M) \neq 0$. If X is orientable, then we can take $M = \mathbf{Z}$ with the trivial structure of Γ-module, otherwise we should take the *orientation module* $\mathbf{Z}_{\mathrm{or}}$. It is isomorphic to $\mathbf{Z}$ as an Abelian group, but $\gamma \in \Gamma$ acts on it by multiplication by -1 if γ reverses the orientation of $X^{\sim}$, and by 1 otherwise. □

This proof at least partially motivates the need for twisted coefficients in the definition of $\mathrm{cd}\,\Gamma$.

As usual, a short exact sequence of coefficients

$$0 \to M' \to M \to M'' \to 0$$

leads to a long exact sequence of cohomology groups:

$$\cdots \to H^n(\Gamma, M') \to H^n(\Gamma, M) \to H^n(\Gamma, M'') \to H^{n+1}(\Gamma, M') \to \cdots.$$

It is the key tool in the proof of the following lemma.

LEMMA 6.1.B. *If* $\mathrm{cd}\,\Gamma < \infty$, *then* $\mathrm{cd}\,\Gamma$ *is equal to the supremum of numbers n such that $H^n(\Gamma, M) \neq 0$ for some* free $\mathbf{Z}[\Gamma]$-*module M.*

PROOF. Let $n = \mathrm{cd}\,\Gamma$. Choose M such that $H^n(\Gamma, M) \neq 0$ and choose a short exact sequence

$$0 \to M' \to F \to M \to 0$$

with F free. Then we have an exact sequence

$$H^n(\Gamma, F) \to H^n(\Gamma, M) \to H^{n+1}(\Gamma, M'),$$

where $H^n(\Gamma, M) \neq 0$ and $H^{n+1}(\Gamma, M') = 0$ because $n + 1 > \mathrm{cd}\,\Gamma$. It follows that $H^n(\Gamma, F) \neq 0$. □

LEMMA 6.1.C. *If X is a finite CW-complex, then $H^n(\Gamma, F) = 0$ for all free $\mathbf{Z}[\Gamma]$-modules F if and only if $H^n(\Gamma, \mathbf{Z}[\Gamma]) = 0$.*

PROOF. The "only if" part is trivial. In order to prove the "if" part, suppose that $H^n(\Gamma, \mathbf{Z}[\Gamma]) = 0$. Then, for any finitely generated free module $F = (\mathbf{Z}[\Gamma])^m$, we have $H^n(\Gamma, F) = H^n(\Gamma, (\mathbf{Z}[\Gamma])^m) = H^n(\Gamma, \mathbf{Z}[\Gamma])^m = 0^m = 0$.

Now, any free module is equal to the direct limit of its finitely generated free submodules. As we will explain in a moment, the fact that X is a finite CW-complex implies that the functor $H^n(\Gamma, \cdot) = H^n(X, \cdot)$ commutes with the direct limits. Hence, the vanishing of the cohomology with coefficients in finitely generated free modules implies the vanishing of the cohomology with coefficients in an arbitrary free module. This completes the proof, modulo the claim about commuting with the direct limits.

It is well known that for any ring R (we are interested in the case $R = \mathbf{Z}[\Gamma]$) and any projective R-module Q the functor $Q \otimes_R \cdot$ commutes with the direct limits. In addition, if P is a *finitely generated* projective R-module, then $\mathrm{Hom}_R(P, M) = P^* \otimes_R M$ for any R-module, where $P^* = \mathrm{Hom}(P, R)$. It follows that for a finitely generated projective R-module P the functor $\mathrm{Hom}_R(P, \cdot)$ commutes with the direct limits.

Returning to our situation, notice that if X is a finite CW-complex and we consider cellular chains, then $C_k(X^\sim)$ are finitely generated free (and hence projective) $\mathbf{Z}[\Gamma]$-modules. Thus the result of the previous paragraph implies that the functor $\mathrm{Hom}_\Gamma(C_k(X^\sim), \cdot)$ commutes with the direct limits. It remains to note that taking the (co)homology groups of a complex always commutes with the direct limits. □

COROLLARY 6.1.D. *If X is a finite CW-complex, then* $\mathrm{cd}\,\Gamma$ *is equal to the maximum of numbers n such that $H^n(\Gamma, \mathbf{Z}[\Gamma]) \neq 0$.*

PROOF. Since X is a finite CW-complex, $\dim X < \infty$. If we use cellular chains, then $C_k(X^\sim) = 0$ for $k > \dim X$, and hence $H^k(\Gamma, M) = H^n(X, M) = 0$ for $k > \dim X$. It follows that $\mathrm{cd}\,\Gamma < \infty$. It remains to apply Lemmas 6.1.B and 6.1.C. □

The last corollary shows the importance of the groups $H^n(\Gamma, \mathbf{Z}[\Gamma])$. Our next goal is to provide a geometric interpretation of these cohomology groups.

LEMMA 6.1.E. *For any Γ-module M the group* $\mathrm{Hom}_\Gamma(M, \mathbf{Z}[\Gamma])$ *is naturally isomorphic to the group* $\mathrm{Hom}_c(M, \mathbf{Z})$ *of all homomorphisms of Abelian groups $f : M \to \mathbf{Z}$ such that for any $m \in M$ the image $f(\gamma m)$ is equal to 0 for almost all $\gamma \in \Gamma$ (i.e., for all γ with the exception of finitely many of them).*

PROOF. Any homomorphism $F : M \to \mathbf{Z}[\Gamma]$ of Γ-modules can be written in the form

$$F(m) = \sum_{\gamma \in \Gamma} f_\gamma(m)\gamma$$

for some homomorphisms $f_\gamma : M \to \mathbf{Z}$. Clearly, for any m the image $f_\gamma(m)$ is equal to 0 for almost all $\gamma \in \Gamma$ (by the definition of $\mathbf{Z}[\Gamma]$). Since F is a homomorphism of Γ-modules, we have $F(\beta m) = \beta F(m)$ for any $m \in M$, $\beta \in \Gamma$. It follows that

$$\sum_{\gamma \in \Gamma} f_\gamma(\beta m)\gamma = \sum_{\gamma \in \Gamma} f_\gamma(m)\beta\gamma,$$

and hence $f_{\beta\gamma}(\beta m) = f_\gamma(m)$ for all $\beta, \gamma \in \Gamma$ and $m \in M$. In particular, $f_\gamma(m) = f_1(\gamma^{-1}m)$ for all $\gamma \in \Gamma$, $m \in M$, where $1 \in \Gamma$ is the unit element of Γ. It follows that F is determined by f_1. Moreover, because for any m the image $f_1(\gamma m) = f_{\gamma\gamma^{-1}}(\gamma m) = f_{\gamma^{-1}}(m)$ is equal to 0 for almost all γ, the homomorphism f_1 belongs to $\mathrm{Hom}_c(M, \mathbf{Z})$.

Conversely, for a homomorphism $f_1 : M \to \mathbf{Z}$ belonging to $\mathrm{Hom}_c(M, \mathbf{Z})$, we can define a homomorphism $F : M \to \mathbf{Z}[\Gamma]$ of Γ-modules by the formula

$$F(m) = \sum_{\gamma\in\Gamma} f_1(\gamma^{-1}m)\gamma.$$

Thus $F \mapsto f_1$ establishes a natural bijection $\mathrm{Hom}_\Gamma(M, \mathbf{Z}[\Gamma]) \to \mathrm{Hom}_c(M, \mathbf{Z})$. □

LEMMA 6.1.F. *If X is a finite CW-complex, then $H^*(X, \mathbf{Z}[\Gamma])$ is naturally isomorphic to the cohomology group with compact support $H_c^*(X^\sim)$ of $X^\sim$.*

PROOF. By Lemma 6.1.E,

$$\begin{aligned} H^*(\Gamma, \mathbf{Z}[\Gamma]) &= H^*(X, \mathbf{Z}[\Gamma]) = H^*(\mathrm{Hom}_\Gamma(C_*(X^\sim), \mathbf{Z}[\Gamma])) \\ &= H^*(\mathrm{Hom}_c(C_*(X^\sim), \mathbf{Z})). \end{aligned}$$

If we use the cellular chains, then $\mathrm{Hom}_c(C_*(X^\sim), \mathbf{Z})$ consists of cochains which are equal to 0 on almost all cells in the preimage of any given cell in X. Because X has only a finite number of cells, these are exactly the cochains equal to 0 on almost all cells in $X^\sim$, i.e., the cochains with compact support. It follows that the last cohomology group is nothing other than the cohomology groups with compact support of $X^\sim$. □

COROLLARY 6.1.G. *If X is a finite CW-complex, then $\mathrm{cd}\,\Gamma$ is equal to the maximum number n such that $H_c^n(X^\sim, \mathbf{Z}) \neq 0$.*

Now, it is the time to make a crucial step and apply the Poincaré–Lefschetz duality.

THEOREM 6.1.H. *Suppose that X is a compact topological manifold of dimension d with boundary (may be, empty) and, simultaneously, is a finite CW-complex. Let m be the minimum number such that $\tilde{H}_m(\partial X^\sim) \neq 0$, where $\tilde{H}_*(\cdot)$, as usual, denotes the* reduced *homology. Then $\mathrm{cd}\,\Gamma = d - m - 1$. (If there is no such m, then $\mathrm{cd}\,\Gamma = 0$.)*

PROOF. By the Poincaré–Lefschetz duality $H_c^n(X^\sim) = H_{d-n}(X^\sim, \partial X^\sim)$. Because X is a $K(\Gamma, 1)$-space, $X^\sim$ is contractible, and hence $H_{d-n}(X^\sim, \partial X^\sim) = \tilde{H}_{d-n-1}(\partial X^\sim)$. It remains to apply Corollary 6.1.G (note that $m = d - n - 1$ if and only if $n = d - m - 1$). □

The assumptions of this theorem are satisfied, for example, when X is a smooth manifold with corners (as in the applications we have in mind) and, hence, can be triangulated.

6.2. *Classifying spaces and universal bundles*

Here we discuss some auxiliary results about classifying spaces and universal bundles which will be needed in 6.3 and 6.4.

Let G be a topological group acting on a topological space X. Let $EG \to BG$ be the Milnor universal principal G-bundle. In particular, $EG \to BG$ is a locally trivial bundle and its total space EG is contractible. We will denote by $EX \to BG$ the associated bundle with the fiber X. Recall that $EX = EG \times_G X$. A choice of a base point $x \in X$ defines a map $G \to X$ given by the formula $g \mapsto gx$ and a map $EG \to EX$ given by the formula $y \mapsto y \times_G x$.

LEMMA 6.2.A. *If $G \to X$ is a Serre fibration, then the map $EG \to EX$ is also a Serre fibration.*

PROOF. Locally, over any sufficiently small open set $U \subset BG$, the bundle $EG \to BG$ is isomorphic to the trivial bundle $U \times G \to U$, the associated bundle is isomorphic to $U \times X \to U$ and the map $EG \to EX$ is the map $U \times G \to U \times X$ induced by $G \to X$. If $G \to X$ is a Serre fibration, then, obviously, $U \times G \to U \times X$ is a Serre fibration also. It follows that $EG \to EX$ is a Serre fibration locally (over EX). By a well known theorem (see, for example, [25, Theorem VII.6.11]), this implies that $EG \to EX$ is a Serre fibration. □

LEMMA 6.2.B. *Let H be the stabilizer of the base point x in G. Suppose that the above map $G \to X$ is a Serre fibration. Then the part*

$$\pi_1(X) \longrightarrow \pi_1(EX) \longrightarrow \pi_1(BG)$$

of the homotopy sequence of the bundle $EX \to BG$ can be canonically identified with the part

$$\pi_1(X) \longrightarrow \pi_0(H) \longrightarrow \pi_0(G)$$

of the homotopy sequence of the fibration $G \to X$.

PROOF. Let $\pi_1(BG) \to \pi_0(G)$ be the boundary map from the homotopy sequence of the bundle $EG \to BG$. Since EG is contractible, it is an isomorphism. By Lemma 6.2.A, the map $EG \to EX$ is a Serre fibration; obviously, its fiber is H. Let $\pi_1(EX) \to \pi_0(H)$ be the boundary map from the homotopy sequence of this fibration. Again, since EG is contractible, it is an isomorphism. Direct check shows that the following diagram

$$\begin{array}{ccccc} \pi_1(X) & \longrightarrow & \pi_1(EX) & \longrightarrow & \pi_1(BG) \\ \downarrow & & \downarrow & & \downarrow \\ \pi_0(X) & \longrightarrow & \pi_0(H) & \longrightarrow & \pi_0(G) \end{array}$$

is commutative, and hence provides the required identification. □

If G acts on another topological space X' and a G-equivariant map $X \to X'$ is given, then there is a natural map $EX \to EX'$ commuting with the projections $EX \to BG$, $EX' \to BG$. If $X \to X'$ is a (weak) homotopy equivalence, then $EX \to EX'$ is also a (weak) homotopy equivalence, and we can identify the homotopy sequences of fibrations $EX \to BG$ and $EX' \to BG$.

6.3. *Mess subgroups*

In this section we discuss some subgroups of the mapping class groups discovered by Mess [163]. These subgroups will be our main tool for estimating the virtual cohomological dimension of the mapping class groups from *below* (cf. the proof of Theorem 6.4.A).

Let S_g be a closed orientable surface of genus g, and let $\mathrm{Mod}_g = \mathrm{Mod}_{S_g}$. Let S_g^1 be an orientable surface of genus g with 1 boundary component, and let $\mathcal{M}_g^1 = \mathcal{M}_{S_g^1}$. This section is devoted to some subgroups B_g, B_g^1 of Mod_g, $\mathcal{M}_g^1$ respectively, where $g \geqslant 2$. They were introduced by Mess [163], and we call them *Mess subgroups*. Although the construction of these involves some choices, it is essentially canonical.

The construction of the Mess subgroups has a recursive character. We start with any subgroup B_2 of Mod_2 generated by Dehn twists about any three pairwise disjoint pairwise nonisotopic circles on S_2. Thus B_2 is isomorphic to $\mathbf{Z}^3$. Suppose that B_g, $g \geqslant 2$, is already defined.

We may assume that S_g is obtained from S_g^1 by gluing a disc D^2 to the boundary of S_g^1. The extension of diffeomorphisms of S_g^1 fixed on ∂S_g^1 by the identity across the disc D^2 defines a homomorphism $\mathcal{M}_g^1 \to \mathrm{Mod}_g$ as in 2.8. Let B_g^1 be the preimage of B_g under this homomorphism.

In order to define B_{g+1}, consider some embedding $S_g^1 \to S_{g+1}$ and identify S_g^1 with its image. The extension of diffeomorphisms of S_g^1 fixed on ∂S_g^1 by the identity across the complement of S_g^1 in S_{g+1} defines a homomorphism $i : \mathcal{M}_g^1 \to \mathrm{Mod}_{g+1}$. Note that the closure of the complement of S_g^1 in S_{g+1} is a torus with one hole. Let us choose some nontrivial circle in this torus with one hole and consider the Dehn twist $t \in \mathrm{Mod}_{g+1}$ about this circle. Let T be the infinite cyclic group generated by t. We define B_{g+1} as the group generated by $i(B_g^1)$ and T.

This completes the construction of the Mess subgroups. A couple of remarks are in order. First, recall (see 2.6) that the restriction of diffeomorphisms of S_g on D^2 defines a fibration $\mathrm{Diff}(S_g) \to \mathrm{Emb}(D^2, S_g)$. As explained in 2.6, its fiber can be identified with $\mathrm{Diff}(S_g^1 \operatorname{fix} \partial)$. Since $\mathrm{Diff}(S_g)$ is simply connected (by Theorem 2.7.G) and $\mathrm{Emb}(D^2, S_g)$ is connected, the homotopy sequence of this fibration ends with

$$1 \longrightarrow \pi_1(\mathrm{Emb}(D^2, S_g)) \longrightarrow \pi_0(\mathrm{Diff}(S_g^1 \operatorname{fix} \partial)) \longrightarrow \pi_0(\mathrm{Diff}(S_g)) \longrightarrow 1,$$

i.e., with

$$1 \longrightarrow \pi_1(\mathrm{Emb}(D^2, S_g)) \longrightarrow \mathcal{M}_g^1 \longrightarrow \mathrm{Mod}_g \longrightarrow 1,$$

where the homomorphism $\mathcal{M}_g^1 \to \mathrm{Mod}_g$ is exactly the one used above. Hence, we have a short exact sequence

$$1 \longrightarrow \pi_1\big(\mathrm{Emb}(D^2, S_g)\big) \longrightarrow B_g^1 \longrightarrow B_g \longrightarrow 1.$$

Since $\mathrm{Emb}(D^2, S_g)$ is weakly homotopy equivalent to the unit tangent bundle $UT(S_g)$ (by Theorem 2.6.C), we can also write this short exact sequence as

$$1 \longrightarrow \pi_1\big(UT(S_g)\big) \longrightarrow B_g^1 \longrightarrow B_g \longrightarrow 1.$$

Note also that the homomorphism $i : \mathcal{M}_g^1 \to \mathrm{Mod}_{g+1}$ is injective by Theorem 2.7.I and that the twist t obviously commutes with the image of i. It follows that B_{g+1} is isomorphic to $B_g^1 \times \mathbf{Z}$. Now we come to the main property of the Mess subgroups (cf. [163, Proposition 1]).

THEOREM 6.3.A. *There exists a closed topological manifold of dimension* $4g - 5$ *which is a* $K(B_g, 1)$*-space. Similarly, there exists a closed topological manifold of dimension* $4g - 2$ *which is a* $K(B_g^1, 1)$*-space. These topological manifolds can be chosen to be simultaneously CW-complexes.*

PROOF. The idea is to start with the 3-dimensional torus, which can be taken as a $K(B_2, 1)$-space and to construct a $K(B_g^1, 1)$-space as the total space of a bundle with the fiber $UT(S_g)$ over an already constructed $K(B_g, 1)$-space, and then get a $K(B_{g+1}, 1)$-space by multiplying this total space by the circle S^1. We would like this bundle to realize the sequence $\pi_1(UT(S_g)) \to B_g^1 \to B_g$ as the π_1-part of its homotopy sequence. Note that $UT(S_g)$ is a $K(\pi, 1)$-space, because it is the total space of a circle bundle over S_g, which is a $K(\pi, 1)$-space. Suppose that a $K(B_g, 1)$-space K_g is already constructed.

Let $G = \mathrm{Diff}(S_g)$, $X = \mathrm{Emb}(D^2, S_g)$ with the obvious action of $\mathrm{Diff}(S_g)$, and let X' be the unit tangent bundle $UT(S_g)$ with the following action of $\mathrm{Diff}(S_g)$: a diffeomorphism acts on a unit vector by its differential and the subsequent normalization. Let $X \to X'$ be the map $u : \mathrm{Emb}(D^2, S_g) \to UT(S_g)$ from 2.6. It is obviously G-equivariant. By Theorem 2.6.C, it is a weak homotopy equivalence. Hence the induced map $EX \to EX'$ is also a weak homotopy equivalence. It follows that we can naturally identify the homotopy sequences of bundles $EX \to BG$ and $EX' \to BG$.

Now, since components of $\mathrm{Diff}(S_g)$ are weakly contractible by Theorem 2.7.G, the classifying space $BG = B\,\mathrm{Diff}(S_g)$ is a $K(\mathrm{Mod}_g, 1)$-space. Hence, the inclusion $B_g \to \mathrm{Mod}_g$ leads to a map $K_g \to BG$ from our $K(B_g, 1)$-space K_g to $BG = B\,\mathrm{Diff}(S_g)$. Consider the bundle $K_g^1 \to K_g$ with the fiber $X' = UT(S_g)$ induced from the bundle $EX' \to BG$ by this map. Since both $UT(S_g)$ and K_g are $K(\pi, 1)$-spaces, the homotopy sequence of the induced bundle shows that K_g^1 is also a $K(\pi, 1)$-space. In addition, this homotopy sequence ends with the short exact sequence

$$1 \longrightarrow \pi_1\big(UT(S_g)\big) \longrightarrow \pi_1\big(K_g^1\big) \longrightarrow \pi_1(K_g) \longrightarrow 1.$$

This short exact sequence naturally maps to the end

$$1 \longrightarrow \pi_1\big(UT(S_g)\big) \longrightarrow \pi_1(EX') \longrightarrow \pi_1(BG) \longrightarrow 1$$

(recall $X' = UT(S_g)$) of the homotopy sequence of the bundle $EX' \to BG$. Moreover, $\pi_1(K_g^1)$ is obviously isomorphic to the preimage of $\pi_1(K_g) = B_g$ under the map $\pi_1(EX') \to \pi_1(BG)$. By the previous paragraph, this short exact sequence is naturally isomorphic to the end

$$1 \longrightarrow \pi_1(X) \longrightarrow \pi_1(EX) \longrightarrow \pi_1(BG) \longrightarrow 1$$

of the homotopy sequence of the bundle $EX \to BG$. Now, consider the fibration $G = \mathrm{Diff}(S_g) \to X = \mathrm{Emb}(D^2, S_g)$. Clearly, $H = \mathrm{Diff}(S_g^1 \text{ fix } \partial)$ is the stabilizer of the embedding $D^2 \to S_g$ and, hence, is the fiber of this fibration. By Lemma 6.2.B, the last short exact sequence is naturally isomorphic to the end

$$1 \longrightarrow \pi_1(X) \longrightarrow \pi_0(H) \longrightarrow \pi_0(G) \longrightarrow 1$$

of the homotopy sequence of the fibration $G \to X$. But this short exact sequence is nothing other than our sequence

$$1 \longrightarrow \pi_1\big(\mathrm{Emb}(D^2, S_g)\big) \longrightarrow \mathcal{M}_g^1 \longrightarrow \mathrm{Mod}_g \longrightarrow 1.$$

It follows that $\pi_1(K_g^1)$ is isomorphic to the preimage of B_g under the map $i : \mathcal{M}_g^1 \to \mathrm{Mod}_g$, i.e., to B_g^1. Hence, we can take K_g^1 as our $K(B_g^1, 1)$-space, and then put $K_{g+1} = K_g^1 \times S^1$.

This completes our construction of $K(B_g, 1)$- and $K(B_g^1, 1)$-spaces. Since they are constructed by consecutively taking the total spaces of locally trivial bundles with 3-manifold fibers and multiplying by S^1 (and starting with the 3-dimensional torus), these spaces are, obviously, topological manifolds. With some additional care, one can turn these spaces into CW-complexes (or even smooth manifolds, and then triangulate them). This proves all statements of the theorem with the exception of the correctness of the dimension values.

Clearly, $\dim K_g^1 = \dim K_g + 3$, $\dim K_{g+1} = \dim K_g^1 + 1$ and $\dim K_2 = 3$. Therefore, $\dim K_g = 4g - 5$ and $\dim K_g^1 = 4g - 2$. This completes the proof. □

COROLLARY 6.3.B. $\mathrm{cd}\, B_g = 4g - 5$, $\mathrm{cd}\, B_g^1 = 4g - 2$.

PROOF. Combine Theorem 6.3.A with Theorem 6.1.A. □

6.4. *Virtual cohomological dimension*

One of the main properties of the cohomological dimension is the inequality $\mathrm{cd}\, \Gamma' \leqslant \mathrm{cd}\, \Gamma$, which holds for $\Gamma' \subset \Gamma$. We refer to [26] for a proof and only note that it uses in a crucial way the fact that all *twisted* coefficients are allowed; cf. [26, Chapter VIII, Proposition 2.4].

It is well known that cd $A = \infty$ for a finite cyclic group A; it follows that any group with torsion has infinite cohomological dimension. In order to get an interesting invariant for groups with torsion, we modify the definition in the way suggested by Serre [213].

If Γ is *virtually torsion free*, i.e., contains a subgroup of finite index which is torsion free, then, by a fundamental theorem of Serre [213] (cf. also [26, Chapter VIII, Theorem 3.1] or [11, Section 3]) all torsion free subgroups of finite index of Γ have the same cohomological dimension. The common value of the cohomology dimension of such subgroups is called the *virtual cohomological dimension* of Γ and is denoted by vcd Γ.

Since Mod_S always has torsion, $\mathrm{cd}\,\mathrm{Mod}_S = \infty$ and the interesting invariant is $\mathrm{vcd}\,\mathrm{Mod}_S$. The results of 6.1 and 6.3 allow us to compute it. We start with closed surfaces.

THEOREM 6.4.A. *If S is a closed surface of genus $g \geqslant 2$, then* $\mathrm{vcd}\,\mathrm{Mod}_S = 4g - 5$. *If S is a sphere or a torus, then* $\mathrm{vcd}\,\mathrm{Mod}_S = 0$ *or* 1 *respectively.*

PROOF. If S is a sphere, then $\mathrm{Mod}_S = 1$ and the result is trivial. If S is a torus, then Mod_S is isomorphic to $\mathrm{SL}_2(\mathbf{Z})$. The last group, as is well known, contains a finitely generated free group as a subgroup of finite index. It is easy to see that the cohomological dimension of a free group is 1. Hence, $\mathrm{vcd}\,\mathrm{Mod}_S = 1$ if S is a torus.

In the rest of the proof we assume that S is a closed surface of genus $g \geqslant 2$. Consider a subgroup Γ of finite index in Mod_S acting freely on the Teichmüller space T_S (such subgroups do exist by 5.4). It is sufficient to show that $\mathrm{cd}\,\Gamma = 4g - 5$ for any such Γ (we do not need to know in advance that such a group Γ is torsion free, since $\mathrm{cd}\,\Gamma < \infty$ implies this).

Consider the intersection $\Gamma' = \Gamma \cap B_g$ of Γ with the Mess subgroup B_g. It is obviously of finite index in B_g. Hence, we can take as a $K(\Gamma', 1)$-space a finite sheeted covering space of the $K(B_g, 1)$-space provided by Theorem 6.3.A. Clearly, this $K(\Gamma', 1)$-space is a closed manifold of dimension $4g - 5$. By Theorem 6.1.A, $\mathrm{cd}\,\Gamma' = 4g - 5$. It follows that $\mathrm{cd}\,\Gamma \geqslant 4g - 5$.

In order to prove the opposite inequality, consider the manifold with corners $T_S(\varepsilon)$ from 5.4 and put $X = T_S(\varepsilon)/\Gamma$. Since the action of Γ on T_S is free, X is also a manifold with corners, and since $T_S(\varepsilon)$ is contractible (cf. 5.4), we can take $T_S(\varepsilon)$ as the universal cover $X^\sim$ of X. By 5.4, the boundary $\partial T_S(\varepsilon)$ is homotopy equivalent to (the geometric realization of) the complex of curves $C(S)$. Hence, Theorem 3.2.C implies that $\partial X^\sim = \partial T_S(\varepsilon)$ is $2g - 3$-connected. Now we apply Theorem 6.1.H. Since $\partial X^\sim$ is $2g - 3$-connected, the number m from this theorem is $\geqslant 2g - 2$ (such an m exist, because we already saw that $\mathrm{cd}\,\Gamma \neq 0$). Finally, $d = \dim X^\sim = \dim T_S(\varepsilon) = \dim T_S = 6g - 6$, and hence Theorem 6.1.H implies that $\mathrm{cd}\,\Gamma \leqslant (6g - 6) - (2g - 2) - 1 = 4g - 5$. This completes the proof. □

THEOREM 6.4.B. *If S is a surface of genus $g \geqslant 2$ with $b \geqslant 1$ boundary components, then* $\mathrm{vcd}\,\mathrm{Mod}_S = 4g - 4 + b$.

PROOF. Let $S_{g,b}$ be a surface of genus g with b boundary components, and let $\mathrm{Mod}_{g,b} = \mathrm{Mod}_{S_{g,b}}$. In order to compute $\mathrm{vcd}\,\mathrm{Mod}_{g,b}$, we will construct a subgroup $\Gamma_{g,b}$ of finite index in $\mathrm{Mod}_{g,b}$ and a $K(\Gamma_{g,b}, 1)$-space. Since the dimension of this space will be $< \infty$, and hence $\mathrm{cd}\,\Gamma_{g,b}$ will be $< \infty$, the subgroup $\Gamma_{g,b}$ will be automatically torsion free, and therefore we will have $\operatorname{vcd}\operatorname{Mod}_{g,b} = \operatorname{cd}\Gamma_{g,b}$. The cohomological dimension $\operatorname{cd}\Gamma_{g,b}$ will be computed with the help of Theorem 6.1.H using our $K(\Gamma_{g,b}, 1)$-spaces.

The construction of subgroups $\Gamma_{g,b}$ has a recursive character. We start with any subgroup $\Gamma_{g,0}$ of finite index in $\operatorname{Mod}_{g,0}$ acting freely on the Teichmüller space $T_{S_{g,0}}$. Let $L_{g,0} = T_{S_{g,0}}(\varepsilon)/\Gamma_{g,0}$, where $T_{S_{g,0}}(\varepsilon)$ is the manifold with corners from 5.4; $L_{g,0}$ will be our $K(\Gamma_{g,0}, 1)$-space. Suppose that a subgroup $\Gamma_{g,b}$ of finite index in $\operatorname{Mod}_{g,b}$ and a $K(\Gamma_{g,b}, 1)$-space $L_{g,b}$ are already constructed.

Let us fix a point x in the interior of $S_{g,b}$. Let $* = \{x\}$; it is a 0-dimensional submanifold of $S_{g,b}$. Clearly, we can identify $\operatorname{Emb}^{\diamond}(*, S_{g,b})$ with the interior $\operatorname{int} S_{g,b}$ of $S_{g,b}$ (recall that the embeddings in $\operatorname{Emb}^{\diamond}(*, S_{g,b})$ can be extended to diffeomorphisms of $S_{g,b}$; see 2.6).

Now, consider the group $\operatorname{Mod}_{g,(b,1)} = \pi_0(\operatorname{Diff}(S_{g,b} \operatorname{fix} *))$. In view of the last paragraph of 2.8, the group $\operatorname{Mod}_{g,(b,1)}$ is isomorphic to the subgroup of the group $\operatorname{Mod}_{g,b+1}$ consisting of the isotopy classes of diffeomorphisms $S_{g,b+1} \to S_{g,b+1}$ preserving (setwise) one fixed boundary component. In particular, $\operatorname{Mod}_{g,(b,1)}$ is isomorphic to a subgroup of finite index (actually, of index $b+1$) of $\operatorname{Mod}_{g,b+1}$.

Let $G = \operatorname{Diff}(S_{g,b})$ and let $X = \operatorname{int} S_{g,b}$ with the obvious action of G. The map $G \to X$ given by $g \mapsto gx$ is nothing but the evaluation of diffeomorphisms at x; it can be identified with the map $\operatorname{Diff}(S_{g,b}) \to \operatorname{Emb}^{\diamond}(*, S_{g,b})$ from 2.6. In particular, Theorem 2.6.A implies that $G \to X$ is a Serre fibration. The stabilizer H of the point $x \in X$ (i.e., the fiber of $G \to X$) is equal to $\operatorname{Diff}(S_{g,b} \operatorname{fix} *)$. The homotopy sequence of this fibration ends with

$$1 \longrightarrow \pi_1(\operatorname{int} S_{g,b}) \longrightarrow \pi_0\big(\operatorname{Diff}(S_{g,b} \operatorname{fix} *)\big) \longrightarrow \pi_0\big(\operatorname{Diff}(S_{g,b})\big) \longrightarrow 1,$$

i.e., with

$$1 \longrightarrow \pi_1(\operatorname{Int} S_{g,b}) \longrightarrow \operatorname{Mod}_{g,(b,1)} \longrightarrow \operatorname{Mod}_{g,b} \longrightarrow 1.$$

By Lemma 6.2.B, this short exact sequence can be naturally identified with

$$1 \longrightarrow \pi_1(X) \longrightarrow \pi_1(EX) \longrightarrow \pi_1(BG) \longrightarrow 1$$

(where $X = \operatorname{int} S_{g,b}$).

Next, we would like to replace the open surface $\operatorname{int} S_{g,b}$ by the compact surface $S_{g,b}$. Let X' be the surface $S_{g,b}$ with the natural action of $G = \operatorname{Diff}(S_{g,b})$. The inclusion $X \to X'$ is, obviously, G-equivariant and is a homotopy equivalence. In view of the remarks at the end of 6.2, we can identify the last exact sequence with

$$1 \longrightarrow \pi_1(X') \longrightarrow \pi_1(EX') \longrightarrow \pi_1(BG) \longrightarrow 1$$

(where $X' = S_{g,b}$). We can write this short exact sequence more explicitly as

$$1 \longrightarrow \pi_1(S_{g,b}) \longrightarrow \pi_1(ES_{g,b}) \longrightarrow \pi_1\big(B\operatorname{Diff}(S_{g,b})\big) \longrightarrow 1.$$

Now, as in the proof of Theorem 6.3.A, we use the fact that $B\operatorname{Diff}(S_{g,b})$ is a $K(\operatorname{Mod}_{g,b}, 1)$-space, and hence the inclusion $\Gamma_{g,b} \to \operatorname{Mod}_{g,b}$ leads to a map $L_{g,b} \to$

$\mathrm{Diff}(S_{g,b})$ from ou[illegible] 1)-space $L_{g,b}$ to $B\,\mathrm{Diff}(S_{g,b})$. Consider the bundle $L_{g,b+1} \to L_{g,b}$ with the [illegible]ed from the bundle $EX' = ES_{g,b} \to BG = B\,\mathrm{Diff}(S_{g,b})$ by this map. Since both $S_{g,b}$ and $L_{g,b}$ are $K(\pi, 1)$-spaces, the homotopy sequence of this bundle shows that $L_{g,b+1}$ is also a $K(\pi, 1)$-space. In addition, this homotopy sequence ends with the short exact sequence

$$1 \to \pi_1(S_{g,b}) \to \pi_1(L_{g,b+1}) \to \pi_1(L_{g,b}) \to 1.$$

The latter naturally maps to the last short exact sequence of the previous paragraph, and a trivial diagram chase shows that $\pi_1(L_{g,b+1})$ is isomorphic to the preimage of $\Gamma_{g,b}$ under the map $\pi_1(ES_{g,b}) \to \pi_1(B\,\mathrm{Diff}(S_{g,b}))$. In view of the above, this preimage is naturally isomorphic to the preimage of $\Gamma_{g,b}$ under the map $\mathrm{Mod}_{g,(b,1)} \to \mathrm{Mod}_{g,b}$. We define $\Gamma_{g,b+1}$ to be equal to this preimage (obviously, it is of finite index in $\mathrm{Mod}_{g,b+1}$) and take $L_{g,b+1}$ as our $K(\Gamma_{g,b+1}, 1)$-space.

This completes the construction of our subgroups $\Gamma_{g,b}$ and $K(\Gamma_{g,b}, 1)$-spaces $L_{g,b}$. Since the latter are constructed by consecutively taking the total spaces of locally trivial bundles with surface (in general, with boundary) fibers starting with a manifold with corners, the spaces $L_{g,b}$ are topological manifolds with boundary and we also can turn them into CW-complexes (compare the proof of Theorem 6.3.A). Obviously, $\dim L_{g,b+1} = \dim L_{g,b} + 2$ for all b.

In order to apply Theorem 6.1.H, we need to analyze the universal covers $L^{\sim}_{g,b}$ of manifolds $L_{g,b}$ and their boundaries $\partial L^{\sim}_{g,b}$. If the universal cover $L^{\sim}_{g,b}$ is already constructed, we can construct $L^{\sim}_{g,b+1}$ in two steps. First, take the bundle $M_{g,b} \to L^{\sim}_{g,b}$ with fiber $S_{g,b}$ induced from the bundle $L_{g,b+1} \to L_{g,b}$ by the map $L^{\sim}_{g,b} \to L_{g,b}$. Since the base $L^{\sim}_{g,b}$ of this bundle is contractible, this bundle is actually trivial, and hence $M_{g,b}$ is homeomorphic to $S_{g,b} \times L^{\sim}_{g,b}$. As the second step, we take the universal cover of $M_{g,b}$. This leads to $S^{\sim}_{g,b} \times L^{\sim}_{g,b}$, where $S^{\sim}_{g,b}$ is the universal cover of the surface $S_{g,b}$. Thus, $L^{\sim}_{g,b+1}$ is homeomorphic to $S^{\sim}_{g,b} \times L^{\sim}_{g,b}$, and hence the boundary $\partial L^{\sim}_{g,b+1}$ is homeomorphic to

$$\partial S^{\sim}_{g,b} \times L^{\sim}_{g,b} \cup S^{\sim}_{g,b} \times \partial L^{\sim}_{g,b}.$$

Now, there are two different cases to consider: $b = 0$ and $b \geqslant 1$.

In the first case, $S^{\sim}_{g,0}$ has no boundary and $\partial L^{\sim}_{g,1}$ is homeomorphic to $S^{\sim}_{g,0} \times \partial L^{\sim}_{g,0}$. Since $S^{\sim}_{g,0}$ is contractible, it follows that $\partial L^{\sim}_{g,1}$ is homotopy equivalent to $\partial L^{\sim}_{g,0}$.

In the second case, $\partial S^{\sim}_{g,b}$ is non-empty and actually consists of a countable (infinite) number of components homeomorphic to the real line. Since $S^{\sim}_{g,b}$ is contractible, it follows that the pair $(S^{\sim}_{g,b}, \partial S^{\sim}_{g,b})$ is homotopy equivalent to the pair (CZ, Z), where Z is a discrete space consisting of a countable number of points and CZ is the cone of Z. The last pair, in turn, is homotopy equivalent to the pair $(\mathbf{R}, \mathbf{Z})$. It follows that $\partial L^{\sim}_{g,b+1}$ is homotopy equivalent to

$$\mathbf{Z} \times L^{\sim}_{g,b} \cup \mathbf{R} \times \partial L^{\sim}_{g,b}.$$

Since the universal covers $L^{\sim}_{g,b}$ are contractible, the last space is homotopy equivalent to a bouquet of an infinite number of suspensions $\Sigma \partial L^{\sim}_{g,b}$.

Now, let $d(b) = \dim L_{g,b}$ and let $m(b)$ be the minimum number m such that $\tilde{H}_m(\partial L^{\sim}_{g,b}) \neq 0$ (we assume that the genus g is fixed). As we saw above, $d(b+1) = d(b) + 2$ for all b. The results of the two previous paragraphs imply that $m(1) = m(0)$ and $m(b+1) = m(b) + 1$ for $b \geqslant 1$. Combining these remarks with Theorem 6.1.H, we see that $\operatorname{cd} \Gamma_{g,1} = \operatorname{cd} \Gamma_{g,0} + 2$ and $\operatorname{cd} \Gamma_{g,b+1} = \operatorname{cd} \Gamma_{g,b} + 1$ for $b \geqslant 1$. Since $\operatorname{cd} \Gamma_{g,0} = \operatorname{vcd} \operatorname{Mod}_{g,0} = 4g - 5$ by Theorem 6.4.A, the theorem follows. □

THEOREM 6.4.C. *If S is a sphere with b holes, then* $\operatorname{vcd} \operatorname{Mod}_S = 0$ *for* $b \leqslant 3$ *and* $\operatorname{vcd} \operatorname{Mod}_S = b - 3$ *for* $b \geqslant 3$. *If S is a torus with b holes, then* $\operatorname{vcd} \operatorname{Mod}_S = 1$ *for* $b = 0$ *and* $\operatorname{vcd} \operatorname{Mod}_S = b$ *for* $b \geqslant 1$.

PROOF. The proof is similar to the proof of the previous theorem. Since for a sphere S with $\leqslant 2$ holes or a torus S without holes the components of $\operatorname{Diff}(S)$ are not contractible (and hence $B \operatorname{Diff}(S)$ is not a $K(\operatorname{Mod}_S, 1)$-space), the inductive argument should start with a sphere with 3 holes and a torus with 1 hole. We leave the details to the reader. □

The first nontrivial estimate $\operatorname{vcd} \operatorname{Mod}_S \leqslant 6g - 9$ of $\operatorname{vcd} \operatorname{Mod}_S$ for a closed surface of genus g was proved in [95,96] (the trivial estimate is $\operatorname{vcd} \operatorname{Mod}_S \leqslant \dim T_S = 6g - 6$). It was deduced from the simply connectivity of $C(S)$, which, in turn, was deduced from the simply connectivity of the Hatcher–Thurston complex [91]. Theorems 6.4.A–6.4.C are due to Harer [76,78]. Another proof was provided by the author [98]. The paper [98] contains also a computation of the virtual cohomological dimension of the mapping class groups of nonorientable surfaces. Both papers [78] and [98] actually prove a stronger result than Theorems 6.4.A–6.4.C. Namely, they prove that Mod_S contains a subgroup of finite index which is a *group with duality* of some dimension in the sense of Bieri–Eckmann [13] and compute this dimension. The *dimension* of a group with duality is built into its definition (like the dimension of a smooth manifold or a Poincaré space); it is always $< \infty$ and turns out to be equal to the cohomological dimension. The last property allows the use of this result for the computation of $\operatorname{vcd} \operatorname{Mod}_S$. We refer to [98, Section 6] for an expository account of this stronger result, and to [26, Chapter VIII], or [11] for more details about groups with duality. The proof of the lower estimate for the cohomological dimension in our proof of Theorem 6.4.A follows the ideas of Mess [163], while the proof of the upper estimate (which turned out to be equal to the lower one) follows the standard approach of [95,96,76,78,98]. The above proof of Theorem 6.4.B seems to be new. Both of these proofs avoid the full power of the theory of groups with duality (a fragment of this theory was included in 6.1) and also do not use the complete description of the homotopy type of $C(S)$ (they use only Theorem 3.2.C, and only for closed surfaces).

6.5. *Homology stability*

In its simplest form, the homology stability for the mapping class groups asserts that the homology group $H_n(\operatorname{Mod}_S)$ (with trivial coefficients $\mathbf{Z}$) does not depend on S provided S is closed and the genus g of S is sufficiently large compared to n. The implied restriction on g is usually called *the domain of stability*. The first result of this sort, with the domain of stability $g \geqslant 3n+1$, was proved by Harer [76,77]. Such results are similar to, and motivated by, the classical homology stability theorems in the algebraic K-theory. The simplest of those asserts that the homology group $H_n(\mathrm{SL}_g(\mathbf{Z}))$ does not depend on g provided g is sufficiently large compared to n. The more advanced versions of the homology stability theorems in algebraic K-theory cover some other natural series of groups, like the special linear groups over rings or the symplectic groups, and usually provide an explicit domain of stability (the role of g is played by some natural parameter). A common feature of the proofs of the homology stability theorems in algebraic K-theory is the need to include into the picture a wider class of groups than the original infinite series: this is required by the way the induction is arranged in the proofs. The situation for the mapping class groups is similar: one needs to consider the groups $\mathcal{M}_S$ for non-closed surfaces S, in addition to the groups Mod_S for closed S. The groups $\mathcal{M}_S$ have a natural advantage over our usual version Mod_S, namely the existence of a natural map $\mathcal{M}_R \to \mathcal{M}_S$ when R is a subsurface of S. This map is given by the extension of diffeomorphisms of R fixed on ∂R by the identity to diffeomorphisms of S. In fact, we can completely switch to the groups $\mathcal{M}_S$ in this context, because $\mathcal{M}_S = \mathrm{Mod}_S$ for closed S (and the homology stability does not hold for the groups Mod_S for non-closed surfaces S; for the second homology group this follows from Theorem 6.6.C).

THEOREM 6.5.A. *Let R be a connected subsurface of a connected surface S. Let g_R be the genus of R. The map $H_m(\mathcal{M}_R) \to H_m(\mathcal{M}_S)$ induced by the natural map $\mathcal{M}_R \to \mathcal{M}_S$ is an epimorphism if*

$$g_R \geqslant 2n+1$$

and is an isomorphism if

$$g_R \geqslant 2n+2.$$

If, in addition, S is not closed, then the map $H_m(\mathcal{M}_R) \to H_m(\mathcal{M}_S)$ is an epimorphism if

$$g_R \geqslant 2n$$

and is an isomorphism if

$$g_R \geqslant 2n+1.$$

The proof of Theorem 6.5.A is fairly technical. In the outline, it follows the Maazen–van der Kallen [119] approach to the homology stability theorems in algebraic K-theory. On the geometric side, one of the basic tools is the action of the groups $\mathcal{M}_S$ on some versions of the complexes of curves $C(S)$. The higher connectivity of these versions plays a crucial role. On the algebraic side, the central role is played by some spectral sequences associated with an action of a group on a simplicial complex and a carefully arranged induction on the genus and the number of boundary components of S. For an introduction to the ideas of the proof of Theorem 6.5.A, we refer to [98, Section 7] and to [109, Section 1]. The details omitted in [98] are provided in [104]. The case of closed S requires an additional argument and is dealt with in [109].

COROLLARY 6.5.B. *If S is a closed surface of genus $\geqslant 2n+2$, then up to isomorphism the homology group $H_n(\mathrm{Mod}_S)$ does not depend on S.*

PROOF. First, recall that $\mathrm{Mod}_S = \mathcal{M}_S$ for closed S.

If S, S' are two closed surfaces of different genera, there is no natural homomorphism between the groups $\mathcal{M}_S$, $\mathcal{M}_{S'}$. In order to compare the homology of $\mathcal{M}_S$, $\mathcal{M}_{S'}$, we can choose two diffeomorphic subsurfaces R, R' of S, S' respectively and apply Theorem 6.5.A to the maps $\mathcal{M}_R \to \mathcal{M}_S$, $\mathcal{M}_{R'} \to \mathcal{M}_{S'}$. If we choose subsurfaces R, R' such that their genus is equal to the lesser of the genera of S, S', the corollary will follow immediately. □

The domain of stability of Theorem 6.5.A and Corollary 6.5.B is substantially better than the Harer's original domain of stability $g \geqslant 3n+1$ from [77]. For non-closed surfaces, the domain of stability $g \geqslant 2n+1$ of Theorem 6.5.A is exactly the same as the best known domain of stability for $H_n(\mathrm{SL}_g(\mathbf{Z}))$. On the other hand, if we consider homology with rational (as opposed to the integer) coefficients, then the domain of stability can be further improved. (A similar phenomenon is well known in the algebraic K-theory.)

THEOREM 6.5.C. *Let R be a connected subsurface of a connected surface S with non-empty boundary. Let g_R be the genus of R. The map $H_n(\mathcal{M}_R, \mathbf{Q}) \to H_n(\mathcal{M}_S, \mathbf{Q})$ induced by the natural map $\mathcal{M}_R \to \mathcal{M}_S$ is an isomorphism if*

$$g_R \geqslant \frac{3n}{2}.$$

For odd n, this map is surjective if

$$g_R \geqslant \frac{3n}{2} - 2.$$

This theorem is due to Harer [84], who also proved that the domain of stability of this theorem cannot be improved. The question about the exact domain of stability for integral coefficients remains open.

Finally, we mention two other stability theorems. In [80], Harer proved a homology stability theorem for subgroups of the mapping class groups consisting of the isotopy classes of diffeomorphisms preserving some spin structure on the surface (these subgroups obviously have finite index in the corresponding mapping class groups). The main result of [109] is a homology stability theorem for so-called *twisted* coefficients, i.e., for coefficients in a non-trivial module. Actually, more important than the fact that the coefficients are twisted is that the natural examples of twisted coefficients substantially depend on the surface in question, i.e., are not *constant*. As the simplest (and typical) example one may consider $H_n(\mathcal{M}_S, H_1(S))$, where $\mathcal{M}_S$ acts on $H_1(S)$ in the natural way. In this special case, we have the following result.

THEOREM 6.5.D. *Let R be a connected subsurface of a connected surface S. Let g_R be the genus of R. Suppose that both R and S have exactly* 1 *boundary component. Then the map $H_n(\mathcal{M}_R, H_1(R)) \to H_n(\mathcal{M}_S, H_1(S))$ induced by the natural maps $\mathcal{M}_R \to \mathcal{M}_S$, $H_1(R) \to H_1(S)$ is an epimorphism if*

$$g_R \geqslant 2n + 1,$$

and is an isomorphism if

$$g_R \geqslant 2n + 3.$$

This is a special case of Corollary 4.9 from [109] (the *coefficient system $S \mapsto H_1(S)$* is of *degree* 1 in the sense of [109]). Looijenga [143] provided a new proof of a special case of the stability theorem for twisted coefficients of [109]. Namely, he considers only coefficients which factor through the natural representation $\mathcal{M}_S \to \mathrm{Sp}_g(\mathbf{Z})$, where g is the genus of S, and which are *rational* vector spaces. In this situation he extends the results of [109] to closed surfaces and, more importantly, explicitly computes the stable homology (i.e., the homology in the domain of stability) with such twisted coefficients in terms of the stable homology with trivial coefficients $\mathbf{Q}$. Looijenga's methods are largely algebro-geometric, but he uses also the homology stability theorem for *constant* coefficients.

Surprisingly, homology stability with twisted coefficients does not hold for closed surfaces, even for the first homology group, as the following theorem implies.

THEOREM 6.5.E. *If S is a closed surface of genus $g \geqslant 2$, then $H_1(\mathcal{M}_S, H_1(S)) = \mathbf{Z}/(2g-2)\mathbf{Z}$.*

This result is due to Morita [171] (cf. [171, Corollary 5.4]).

6.6. *The low-dimensional homology groups*

We start by computing the first homology group of PMod_S.

LEMMA 6.6.A. *$H_1(\mathrm{PMod}_S)$ depends only on genus of S.*

PROOF. It is sufficient to show that $H_1(\mathrm{PMod}_S)$ does not change if we make a hole in S. So, let R be the result of removing from S the interior of a disc embedded in S. Consider the exact sequence

$$\pi_1(S) \to \mathrm{PMod}_R \to \mathrm{PMod}_S \to 1$$

from 2.8. If $a \in \pi_1(S)$ is the homotopy class of an embedded loop in S, then the image of a in PMod_R has the form $t_\alpha t_\beta^{-1}$ for some circles α, β in R. Moreover, if a is the homotopy class of an embedded non separating loop, then both circles α and β are non separating; see Lemma 4.1.I. It follows that t_α is conjugate to t_β, and hence their images in $H_1(\mathrm{PMod}_R)$

are equal. Therefore, the image of a in $H_1(\text{PMod}_R)$ is 0. Since $\pi_1(S)$ is generated by the homotopy classes of embedded non separating loops (if S has non-empty boundary, then some of the standard generators are separating, but they can easily be replaced by two non separating generators), the image of $\pi_1(S)$ in $H_1(\text{PMod}_R)$ is 0. Using the above exact sequence, we see that $H_1(\text{PMod}_R)$ is isomorphic to $H_1(\text{PMod}_S)$. □

THEOREM 6.6.B. *Let S be a surface of genus g. The first homology group $H_1(\text{PMod}_S)$ is equal to $\mathbf{Z}/12\mathbf{Z}$ if $g = 1$, to $\mathbf{Z}/10\mathbf{Z}$ if $g = 2$, and to 0 if $g \geqslant 3$.*

PROOF. If S is a closed surface of genus 1 (i.e., a torus), then PMod_S is isomorphic to $\text{SL}_2(\mathbf{Z})$ and the result is well known (to algebraists, at least). If S is a closed surface of genus 2, then, probably, the best way to compute $H_1(\text{PMod}_S)$ is to use the presentation of Mod_S due to Bergau and Mennicke [12] and Birman and Hilden [19]; cf. end of 4.3. Given the presentation, the computation of $H_1(\text{PMod}_S)$ is routine and can be left to the reader. This proves the theorem for closed surfaces of genus 1 and 2, and then Lemma 6.6.A implies it for all surfaces of genus 1 and 2.

In the generic case $g \geqslant 3$ we actually do not need Lemma 6.6.A. In this case, we can embed the sphere with 4 holes S_0^4 from Lemma 4.1.H in S in such a way that all circles C_i, $0 \leqslant i \leqslant 3$, and C_{ij}, $1 \leqslant i < j \leqslant 3$, will be non-separating in S; see Figure 13. For

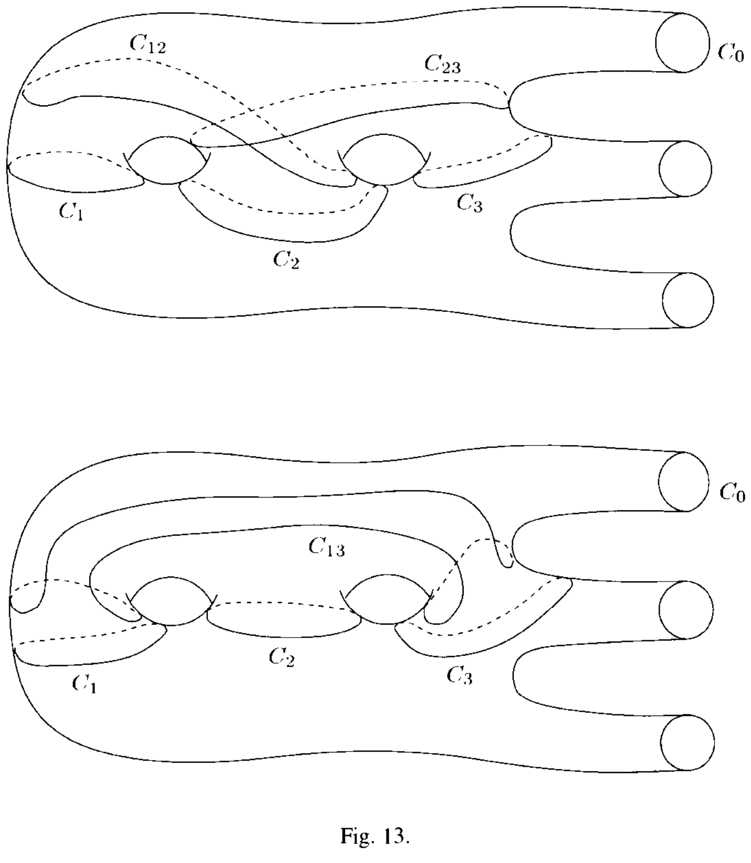


Fig. 13.

such an embedding all Dehn twists t_i, t_{ij} from the lantern relation are conjugate, and hence represent the same element of $H_1(\mathrm{PMod}_S)$. The lantern relation $t_0t_1t_2t_3 = t_{12}t_{13}t_{23}$ implies that this element is equal to 0. On the other hand, by Theorem 4.2.C, PMod_S is generated by Dehn twists about non separating circles. All such Dehn twists are conjugate to t_0, and hence represent 0 in $H_1(\mathrm{Mod}_S)$. It follows that $H_1(\mathrm{PMod}_S) = 0$. □

The path to this nice argument for the genus $\geqslant 3$ case, which is due to Harer [75], was fairly long. First, Mumford [186] proved that $H_1(\mathrm{Mod}_S)$ is cyclic of order dividing 10 for closed surfaces S of genus $\geqslant 2$. Then Birman [15] computed $H_1(\mathrm{Mod}_S)$ for closed surfaces S of genus 1 and 2, and proved that for genus $\geqslant 3$ the order of this group is 1 or 2. Finally, Powell [205] proved that this order is actually 1 by means of some complicated computations.

THEOREM 6.6.C. *Let S be a surface of genus g with b boundary components. If $g \geqslant 5$, then $H_2(\mathrm{PMod}_S) = \mathbf{Z}^{b+1}$.*

This result is due to Harer [75]. One should warn the reader that [75] contains some mistakes, even in the statement of the main theorem (it is asserted in [75] that $H_2(\mathrm{PMod}_S) = \mathbf{Z} \oplus \mathbf{Z}/(2g-2)\mathbf{Z}$ for closed surfaces S of genus g). But the general agreement seems to be that Theorem 6.6.C is correct. The main tool in [75] is an action of PMod_S on a modification of the Hatcher–Thurston complex [91] and the spectral sequences associated with this action. Later on, Harer [81] suggested an alternative method to compute the second homology group with *rational* coefficients, and also extended this new method to the third homology group. The approach of [81] is close in spirit to the proofs of the homology stability theorems 6.5, and uses the actions of the mapping class groups on complexes similar to those used to prove the stability theorems. The spectral sequences associated with these actions relate the homology groups of a mapping class group with the homology groups of the stabilizers of the simplices, which are, in fact, mapping class groups of smaller surfaces. This allows use of an inductive process for the computations. Let us state the main result of [81].

THEOREM 6.6.D. *Let S be a surface of genus g.*
$H_2(\mathcal{M}_S, \mathbf{Q})$ is equal to 0 if $g = 2$, and to $\mathbf{Q}$ if $g \geqslant 3$.
$H_3(\mathcal{M}_S, \mathbf{Q}) = 0$ if $g \geqslant 6$.

Using the same approach, in [82] Harer computed $H_2(\cdot, \mathbf{Q})$ for subgroups of the mapping class groups consisting of the isotopy classes of diffeomorphisms preserving some spin structure on the surface, and in [83] he also computed the fourth homology group of PMod_S. More precisely, the main result of [83] is the following theorem.

THEOREM 6.6.E. *Let S be a surface of genus g with b boundary components.*
$H_4(\mathcal{M}_S, \mathbf{Q}) = \mathbf{Q}^2$ if $g \geqslant 6$, $b \geqslant 1$ or if $g \geqslant 10$.

The computations involved in the proof of this theorem are extremely complicated. Finally, in [72] Hain and Looijenga report that recently Harer has extended his computations to $H_5(\mathrm{PMod}_S, \mathbf{Q})$. This homology group turns out to be 0 for sufficiently high genus.

6.7. *Virtual Euler characteristic*

In the spirit of Serre's approach [213], the usual Euler characteristic $\mathrm{ch}(\cdot) = \sum_{i \geqslant 0} \operatorname{rank} H^i(\cdot)$ is not the right invariant for discrete groups Γ with torsion, even when it is well defined. One should consider instead the so-called *virtual Euler characteristic* $\chi(\cdot)$. If Γ contains a subgroup Γ' of finite index admitting a finite CW-complex as a $K(\Gamma', 1)$-space (this implies that Γ' is torsion free), then, by definition,

$$\chi(\Gamma) = \frac{1}{[\Gamma : \Gamma']} \mathrm{ch}(\Gamma').$$

It turns out that this number does not depend on the choice of Γ'. In general and in the most interesting cases, $\chi(\Gamma)$ is not an integer, but only a rational number.

For the virtual Euler characteristic there is an analogue of the usual formula computing the Euler characteristic of a CW-complex by counting the number of cells of various dimensions. Suppose that Γ acts on a CW-complex T preserving its CW-structure. Suppose that, in addition, T is contractible, some subgroup Γ' of finite index acts freely on T (and hence T/Γ' is a $K(\Gamma', 1)$-space), the number of orbits of cells is finite and the isotropy groups of cells are finite. Then

$$\chi(\Gamma) = \sum_{\sigma} (-1)^{\dim \sigma} \frac{1}{[\Gamma_\sigma : 1]},$$

where Γ_σ is the isotropy group of the cell σ, $[\Gamma_\sigma : 1]$ is its order and σ runs over some set of representatives of the orbits of cells. If Γ acts on T freely, this formula reduces to the usual formula for the Euler characteristic of the $K(\Gamma, 1)$-space T/Γ. For more details about the virtual Euler characteristic we refer to [213] and to [26, Section IX.7].

THEOREM 6.7.A. *If S has one boundary component, then $\chi(\mathrm{Mod}_S) = -B_{2g}/2g = \zeta(1-2g)$, where g is the genus of S, B_{2g} is the $2g$-th Bernoulli number and ζ is the Riemann ζ-function.*

PROOF. Let R be a once-punctured surface of the same genus as S. By 5.1, Mod_S is isomorphic to Mod_R. Hence, we may consider Mod_R instead of Mod_S.

We will use the ideal triangulation of the Teichmüller space T_R introduced in 5.5. Let us consider the (first) barycentric subdivision of the complex $A(R)$, and take the union $\Sigma(R)$ of the barycentric stars of those simplices of $A(R)$ which are *not* simplices of $A_\infty(R)$. This union has a natural structure of a CW-complex with the barycentric stars as cells. One may say that $\Sigma(R)$ is the complex dual to the ideal triangulation of T_R from 5.5. It is not hard to see that $\Sigma(R)$ is a Mod_R-equivariant deformation retract of T_R. In particular, $\Sigma(R)$ is contractible. Cells of $\Sigma(R)$ are in one-to-one correspondence with simplices of $A(R)$ which are not simplices of $A_\infty(R)$, i.e., with ideal cell decompositions of R. Under this correspondence, the orbits of cells correspond to the *topological types* of ideal cell decompositions. Since the number of these topological types is obviously finite, the number of orbits is also finite. The isotropy group of a cell corresponding to an ideal cell decomposition is, clearly, equal to the group of isotopy classes of diffeomorphisms preserving this ideal cell decomposition, and hence is finite. (On the contrary, the isotropy groups of simplices in $A_\infty(R)$ are infinite.) As in the case of Teichmüller spaces T_S of surfaces with boundary (see 5.4), there are subgroups of finite index in Mod_R acting freely on T_R, and hence on $\Sigma(R)$. It follows that we can use the CW-complex $\Sigma(R)$ with its natural action of Mod_R in order to compute $\chi(\mathrm{Mod}_R)$ by the above formula.

The application of this formula to this situation is not a simple matter at all. The problem is essentially a combinatorial one, since an ideal cell decomposition of R is an essentially combinatorial object, and the corresponding isotropy group is the group of symmetries of this combinatorial object. But there is no simple way to list all ideal cell decompositions (up to homeomorphism, of course) and to find the orders of their groups of symmetries. By some breathtaking combinatorial arguments, looking like sort of a miracle for an uninitiated topologist or algebraic geometer, one can compute exactly the combination of these orders leading to $\chi(\mathrm{Mod}_R)$. We refer to the original papers of Harer and Zagier [85] and Penner [199] for the details. While in [85] the needed technique (integration over the space of Hermitian matrices!) appears as a *deus ex machine*, Penner [199] motivates it by some ideas from quantum physics. A simplified proof was provided by Kontsevich [126]; see [126, Appendix D]. More recently, Zagier [243] provided a considerably simpler proof of the key combinatorial fact needed for the calculation of $\chi(\mathrm{Mod}_R)$. This new proof is based on the theory of characters of symmetric groups (and avoids the integration over the space of Hermitian matrices). Despite this dramatic simplification, the original proofs retain, in the author's view, considerable interest. □

COROLLARY 6.7.B. *If S is a closed surface of genus $g \geqslant 2$, then $\chi(\mathrm{Mod}_S) = B_{2g}/4g(g-1)$.*

PROOF. The corollary immediately follows from the theorem if we use the short exact sequence

$$1 \to \pi_1(S) \to \mathrm{Mod}_{S_1} \to \mathrm{Mod}_S \to 1,$$

where S_1 is the result of making one hole in S (cf. 2.8), the fact that virtual Euler characteristic is multiplicative in short exact sequences (cf. [26, Chapter IX, Proposition 7.3]), and the easy fact that $\chi(\pi_1(S))$ is equal to the Euler characteristic of S, i.e., to $2-2g$. □

Finally, we note that [85] also contains some information about the usual Euler characteristic $\sum(-1)^i \dim H_i(\mathrm{Mod}_S, \mathbf{Q})$.

6.8. *Torsion in mapping class groups, torsion in their cohomology and related topics*

Mapping class groups contain many torsion elements, which can be exploited to detect torsion in their cohomology. Clearly, any finite group Γ acting on a surface S by orientation-preserving diffeomorphisms may be considered as a finite subgroup of $\mathrm{Diff}(S)$, and hence leads to a finite subgroup of Mod_S. In fact, the latter subgroup of Mod_S is isomorphic to Γ, as immediately follows from the next theorem.

THEOREM 6.8.A. *Let S be an orientable surface of negative Euler characteristic. If Γ is a finite subgroup of* $\mathrm{Diff}(S)$*, then the canonical homomorphism*

$$\Gamma \to \mathrm{Aut}\big(H_1(S, \mathbf{Z}/m\mathbf{Z})\big)$$

defined by the action of diffeomorphisms on homology is injective for any $m \geqslant 3$.

Note that the homomorphism $\Gamma \to \mathrm{Aut}(H_1(S, \mathbf{Z}/m\mathbf{Z}))$ obviously factors through Mod_S. This theorem is essentially due to Serre [212]; for an elementary proof, see [108, Theorem 1.3 and Supplement 1.4].

According to a famous theorem of Kerckhoff [122,123], every finite subgroup of Mod_S can be obtained as the image of a finite subgroup of $\mathrm{Diff}(S)$; see Theorem 7.4.B. In practice finite subgroups of Mod_S are always obtained from an action of a finite group on S, i.e., as images of finite subgroups of $\mathrm{Diff}(S)$. Kerckhoff's theorem tells us that we will not miss any finite subgroup by restricting ourselves to this method. Note that every finite group admits an embedding into some mapping class group because, as is well known, every finite group occurs as a group of symmetries of a Riemann surface. Moreover, Greenberg [66] proved that every finite group is isomorphic to the group of all automorphisms of some Riemann surface.

Finite subgroups of Mod_S are used to detect nontrivial cohomology classes of Mod_S by restricting cohomology classes to appropriate finite subgroups. Of course, one can detect in this way only torsion cohomology classes, because cohomology groups of finite groups always consist of torsion classes. The first result of this sort was obtained by Charney and Lee [35]. It was improved by Glover and Mislin [62,63] and by Charney and Lee [36]. It is convenient to state these results in terms of the stable cohomology groups in the sense of homology stability discussed in 6.5. Note that there are cohomology stability theorems exactly similar to the homology stability theorems of 6.5. In fact, stability theorems for cohomology follow from the corresponding stability theorems for homology. In particular, $H^m(\mathcal{M}_S)$ does not depends on S if the genus of S is sufficiently large compared to m. Let us denote by $H^m(\mathcal{M})$ the group $H^m(\mathcal{M}_S)$ for surfaces S of sufficiently large genus. Similar notation will be used for (untwisted) coefficients other than $\mathbf{Z}$.

THEOREM 6.8.B. *For all k the stable cohomology group $H^{4k}(\mathcal{M})$ contains an element of order equal to the denominator of the rational number $B_{2k}/2k$, where B_{2k} is the $2k$-th Bernoulli number.*

This theorem is due to Glover and Mislin [62,63]. The numbers appearing in this theorem are the same as in the computation of the virtual Euler characteristic (cf. Theorem 6.7.A). This remarkable fact apparently has no (known) explanation. In order to construct elements of $H^*(\mathcal{M})$, Glover and Mislin consider closed surfaces S and use the canonical representation $\mathcal{M}_S \to \mathrm{Aut}(H_1(S, \mathbf{C}))$ defined by the action of diffeomorphisms on homology. This leads to flat complex vector bundle (with the fiber $H_1(S, \mathbf{C})$)

over $K(\mathcal{M}_S, 1)$, and the Chern classes of this bundle can be considered as elements of $H^*(\mathcal{M}_S)$. Since the bundle is flat, they are torsion elements. One needs to know that these Chern classes are non-zero and, moreover, have sufficiently high order. This is checked by restricting (pulling back) them to specially constructed finite cyclic subgroups of $\mathcal{M}_S$. The elements alluded to in the theorem are some combinations of these Chern classes.

The cyclic subgroups used to detect the order of cohomology classes from Theorem 6.8.B are constructed only for surfaces of fairly high genus (compared to the dimension of the classes). The (co)homology stability theorems imply that these cohomology classes survive till a much lower genus. It happens that sometimes these cyclic groups cannot act nontrivially on surfaces of such a low genus, and hence cannot be subgroups of the mapping class groups of these low genus surfaces. In this way, one gets examples of p-torsion cohomology classes of mapping class groups having no p-torsion elements (where p is some prime number). Such torsion elements in cohomology are somewhat unusual and are called *exotic* torsion elements. The existence of exotic torsion elements was observed in [63].

THEOREM 6.8.C. *The stable cohomology* $H^*(\mathcal{M}, \mathbf{Z}[1/2])$ *with coefficients* $\mathbf{Z}[1/2]$ *contains as a direct summand the cohomology* $H^*(\operatorname{Im} J, \mathbf{Z}[1/2])$ *of the space* $\operatorname{Im} J$ *with the same coefficients.*

This theorem is due to Charney and Lee [36]. The space $\operatorname{Im} J$ is a well known character from homotopy theory. It is known that its cohomology groups consist of torsion elements; replacing the coefficients $\mathbf{Z}$ by $\mathbf{Z}[1/2]$ just kills the 2-torsion. Charney and Lee use some cyclic subgroups in a manner similar to that of Glover and Mislin [63] (but the subgroups of Charney and Lee are smaller and simpler to construct) together with some deep ideas from homotopy theory and algebraic K-theory (the methods of Glover and Mislin are more elementary). Another difference is that Charney and Lee use the canonical homomorphisms $\mathcal{M}_S \to \operatorname{Aut}(H_1(S, \mathbf{Z}/p\mathbf{Z}))$ (their images are contained in the symplectic groups over the field $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$, and a lot is known about the cohomology of these groups from the work of Quillen and others) instead of the homomorphisms $\mathcal{M}_S \to \operatorname{Aut}(H_1(S, \mathbf{C}))$ used by Glover and Mislin. Theorem 6.8.C can be used to produce some torsion classes in addition to those constructed by Glover and Mislin [63].

One can easily realize $H^*(\mathcal{M})$ as the cohomology of some space. Let S_g^1 be a surface of genus g with one boundary component. We may assume that S_g^1 is contained in S_{g+1}^1 (and hence $S_{g+1}^1 \setminus \operatorname{int} S_g^1$ is a torus with two holes) for all $g \geqslant 1$. Then we have a sequence of inclusions $S_1^1 \to \cdots \to S_g^1 \to S_{g+1}^1 \to \cdots$, inducing inclusions of the corresponding mapping class groups $\mathcal{M}_1^1 \to \cdots \to \mathcal{M}_g^1 \to \mathcal{M}_{g+1}^1 \to \cdots$, where $\mathcal{M}_n^1 = \mathcal{M}_{S_n^1}$. Let $\mathcal{M}$ be the direct limit (i.e., the union) of these groups. Clearly, the cohomology $H^*(\mathcal{M})$ of this group is exactly what was denoted by $H^*(\mathcal{M})$ above. By the definition, $H^*(\mathcal{M})$ is the cohomology $H^*(K(\mathcal{M}, 1))$ of the space $K(\mathcal{M}, 1)$. A natural problem is to refine Theorem 6.8.C and produce a splitting on the level of spaces, not just cohomology groups. In fact, it is natural to replace $K(\mathcal{M}, 1)$ by the simply connected space $K(\mathcal{M}, 1)^+$ having the same cohomology groups as $K(\mathcal{M}, 1)$. Here $^+$ denotes Quillen's plus-construction, which is well-defined for $K(\mathcal{M}, 1)$ because $H_1(K(\mathcal{M}, 1)) = 0$ as a corollary of Theorem 6.6.B.

Charney and Cohen [34] proved that it is possible to find such a splitting in the stable homotopy category. Note that the difference between $K(\mathcal{M}, 1)$ and $K(\mathcal{M}, 1)^+$ disappears in the stable homotopy category.

THEOREM 6.8.D. $(\mathrm{Im}\, J) \otimes \mathbf{Z}[1/2]$, *i.e., the space* $\mathrm{Im}\, J$ *localized away from* 2, *is a stable retract of* $K(\mathcal{M}, 1)$.

The proof uses some noncyclic finite groups π_k (they are iterated wreath products of cyclic groups). But Charney and Cohen do not embed them in any $\mathcal{M}_S$. Instead of this, they construct maps $K(\pi_k, 1) \to K(\mathcal{M}, 1)^+$ (in fact, only over finite skeletons) using the homology stability theorems for mapping class groups in order to split natural maps of the form $K(\mathcal{M}_S, 1)^+ \to K(\mathcal{M}_R, 1)^+$, where S is a surface with one boundary component and R is the result of gluing a disc to this boundary component.

Recently, Tillmann [225] proved that the splitting of Theorem 6.8.C indeed can be realized on the level of spaces without passing to the stable homotopy category. Namely, she proved the following theorem.

THEOREM 6.8.E. *There is a space* Y *such that* $K(\mathcal{M}, 1)^+$ *is homotopy equivalent to* $((\mathrm{Im}\, J) \otimes \mathbf{Z}[1/2]) \times Y$.

The proof is based on Theorem 6.8.D and the following fundamental theorem of Tillmann [225].

THEOREM 6.8.F. $K(\mathcal{M}, 1)^+$ *has the homotopy type of an infinite loop space.*

This theorem is obviously of great interest independent of its applications to Theorem 6.8.E. The proof is based on a construction of suitable categories (using the results of a previous paper of Tillmann [224]) which lead to an acceptable input for an infinite loop space machine. Then a new group completion theorem and Harer's homology stability theorem (see 6.5) are used to identify the resulting infinite loop space with $K(\mathcal{M}, 1)^+$. Tillmann's infinite loop space structure on $K(\mathcal{M}, 1)^+$ extends the loop space structure on $K(\mathcal{M}, 1)^+$ constructed previously by more elementary means by Miller [164] and Böedigheimer [22]. But it is not clear yet if it extends the double loop space structure of [164] and [22]. See also [42] for related results about the homology operations in $H^*(\mathcal{M})$.

Now we turn to the mapping class group $\mathrm{Mod}_2 = \mathrm{Mod}_{S_2}$ of a closed surface S_2 of genus 2. Lee and Weintraub [131] and Cohen [37] proved that $H^*(\mathrm{Mod}_2, \mathbf{Z})$ is a torsion group, having nontrivial p-torsion (for prime p) only for $p = 2, 3$ or 5. Cohen [37,38] showed that the 5-torsion is completely captured by a cyclic subgroup of order 5. More precisely, the following theorem holds.

THEOREM 6.8.G. *There is a homomorphism* $\mathbf{Z}/5\mathbf{Z} \to \mathrm{Mod}_2$ *inducing an isomorphism on* 5*-primary components in homology and on homology with coefficients* $\mathbf{Z}/5\mathbf{Z}$*. In particular,* $H_n(\mathrm{Mod}_2, \mathbf{Z})$ *is equal to* $\mathbf{Z}/5\mathbf{Z}$ *plus some* 2*- and* 3*-torsion groups for odd* n *and is a sum of* 2*- and* 3*-torsion groups for even* n.

The cohomology of Mod_2 with coefficients $\mathbf{Z}/2\mathbf{Z}$, $\mathbf{Z}/3\mathbf{Z}$, and $\mathbf{Z}/5\mathbf{Z}$ was essentially completely described by Cohen and Benson [38,9,10]. For the coefficients $\mathbf{Z}/3\mathbf{Z}$ and $\mathbf{Z}/5\mathbf{Z}$, even the ring structure of the cohomology was determined, while for the coefficients $\mathbf{Z}/2\mathbf{Z}$ the ranks of the cohomology groups were computed, but the question of which of two possible ring structures is correct was left open. These papers also contain some information about the integral cohomology of Mod_2, the structure of which was further elucidated by Cohen [39]. We state only the result for the coefficients $\mathbf{Z}[1/2]$, which is especially simple.

THEOREM 6.8.H. *The ring* $H^*(\mathrm{Mod}_2, \mathbf{Z}[1/2])$ *is isomorphic to*

$$\mathbf{Z}[1/2][x, y, z]/(5x, 3y, 3z, z^2),$$

where x, y, z *have degrees* 2, 4, 5 *respectively.*

The proof invokes a connection between Mod_2 and the mapping class group $\mathrm{Mod}_0^6 = \mathrm{Mod}_{S_0^6}$ of a sphere with six holes S_0^6 due to Birman and Hilden [19]. Namely, there is a short exact sequence

$$1 \to \mathbf{Z}/2\mathbf{Z} \to \mathrm{Mod}_2 \to \mathrm{Mod}_0^6 \to 1,$$

with the image of $\mathbf{Z}/2\mathbf{Z}$ generated by the hyperelliptic involution (cf. end of 4.3 for the latter). In addition to this, a close relation between Mod_0^6 and the Artin braid group on six strings is exploited. This geometric input leads to cohomological computations by means of some powerful homotopy-theoretic techniques. For further results obtained by this approach we refer to [40,41], where some noncyclic finite subgroups of $\mathcal{M}_S$, namely quaternion and dihedral groups in $\mathcal{M}_S$, play an essential role.

For further information about many of the topics discussed in this section, and for other related results, we refer to an excellent expository paper [167] by Mislin. In particular, it contains a discussion of results of Glover, Mislin and Xia [64,65], and of related results of Xia [235–239]; see also [240–242] and [207] for further results.

7. Thurston's theory and its applications

This section is devoted to an overview of Thurston's theory of surface diffeomorphisms and some of its applications. For more details, we refer to Thurston's updated announcement of the main results [222], and for a detailed exposition to the proceedings of the Orsay seminar [53]. In the case of closed surfaces, a nice introduction from the point of view of hyperbolic geometry is provided by Casson and Bleiler [30]. See also Hatcher [90].

7.1. *Classification of mapping classes*

An element $f \in \mathrm{Mod}_S$ is called *reducible* if it fixes some simplex of $C(S)$ (but, perhaps, permutes its vertices), and *irreducible* otherwise. As usual, we say that an element f is of *finite order* if $f^n = 1$ for some $n \neq 0$. If an element f is not of finite order and is irreducible, then it is called *pseudo-Anosov*. It turns out that elements in all three classes (reducible, finite order, pseudo-Anosov) can be represented by diffeomorphisms having some special properties. These special representatives play a crucial role in the understanding of corresponding mapping classes.

First of all, for a reducible element f fixing a simplex σ of $C(S)$, let us consider the union C of some disjoint circles representing the vertices of σ. We will call such a union a *realization* of σ. In this case there is a diffeomorphism $F: S \to S$ representing the isotopy class f and leaving C invariant; $F(C) = C$. Such a diffeomorphism F induces a diffeomorphism $F_C: S_C \to S_C$ of the result S_C of cutting S along C. The components of S_C are, in a definite sense, simpler than S (they have larger Euler characteristic) and this allows the use this procedure in inductive arguments. One minor difficulty comes from the fact that S_C has, in general, several components, and F_C may permute them in a nontrivial way. The standard way to deal with this is to replace F_C by some power of it not permuting the components. Another, more serious, problem comes from the fact that F_C may be isotopic to the identity, and then we lose all the information after the cutting. This happens exactly when f is a product of Dehn twists around the components of C. This case needs to be dealt with separately. Anyhow, any reducible mapping class can be represented by a *reducible* diffeomorphism, i.e., a diffeomorphism preserving the union of several pairwise disjoint and non-isotopic nontrivial circles.

While the results outlined in the previous paragraph are essentially elementary, the next case, namely the case of elements of finite order, is much deeper. Special representatives of such elements are provided by the following theorem of Nielsen [194].

THEOREM 7.1.A. *If $f \in \mathrm{Mod}_S$ is an element of finite order $n \neq 0$, then f can represented by a diffeomorphism $F: S \to S$ of the same order n. Moreover, one can choose F to be an isometry of a metric of constant curvature on S with geodesic boundary.*

See [53, Exposé 11, §4] for a proof of this theorem in the context of Thurston's theory and an outline of another proof due to Fenchel [54]. As a simple application of this theorem, consider the kernel $\Gamma_S(m)$ of the natural homomorphism

$$\mathrm{Mod}_S \to \mathrm{Aut}\big(H_1(S, \mathbf{Z}/m\mathbf{Z})\big)$$

defined by the action of diffeomorphisms on homology. Clearly, $\Gamma_S(m)$ has finite index in Mod_S.

COROLLARY 7.1.B. *If $m \geqslant 3$, then $\Gamma_S(m)$ is torsion free.*

PROOF. If the Euler characteristic of S is negative, it is sufficient to combine Nielsen's Theorem 7.1.A with Serre's Theorem 6.8.A. The remaining cases are elementary and left to the reader. □

Finally, the remaining elements, which we called pseudo-Anosov, also can be represented by some very special diffeomorphisms, namely by *pseudo-Anosov diffeomorphisms*.

(To be honest, they are only homeomorphisms, but their non-smoothness is very mild and harmless.) We will not give their definition here, which generalizes (the two-dimensional case of) the definition of *Anosov diffeomorphisms*, introduced by Anosov [2] and playing a central role in the theory of dynamical systems; see [53] or [222]. The existence of pseudo-Anosov diffeomorphisms is one of the main achievements of Thurston's theory of the diffeomorphisms of surfaces.

While the existence of nontrivial reducible mapping classes or mapping classes of finite order is quite obvious, it is not so for pseudo-Anosov mapping classes. Let us indicate a couple of ways to get such mapping classes. We need the following notion: we say that a set A of vertices of $C(S)$ (i.e., a set of isotopy classes of nontrivial circles on S) *fills* S if for every vertex γ there is a vertex $\alpha \in A$ such that any two circles representing α, γ have a non-empty intersection. For finite sets A, this is equivalent to the following property: for any collection of circles representing all isotopy classes from A, all components of the complement of the union of these circles in S are either (open) discs or (half-open) annuli, and each of the latter contains a component of ∂S.

THEOREM 7.1.C. *If α, β are two isotopy classes of nontrivial circles such that $\{\alpha, \beta\}$ fills S, then all elements $t_\alpha^n t_\beta^{-m}$ for $m, n \geqslant 1$ are pseudo-Anosov.*

See [53], Exposé 13, Remark 2 after Theorem III.3. Theorem III.3 itself shows that under the assumptions of Theorem 7.1.C, many other combinations of t_α, t_β are also pseudo-Anosov. These results were generalized by Long [138] and then by Penner [200] to the situations when more than two Dehn twists are involved. Further examples were provided by Bauer [7]. The following remarkable result, due to Fathi [52], while not giving completely explicit examples of pseudo-Anosov elements, shows that they are in a very strong sense generic.

THEOREM 7.1.D. *Let $f \in \mathrm{Mod}_S$ and α be a vertex of $C(S)$. If the set $\{f^n(\alpha)\colon n \in \mathbf{Z}\}$ fills S, then all elements $t_\alpha^n f$ are pseudo-Anosov except for at most 7 consecutive values of n.*

For example, if $\{\alpha, \beta\}$ fills S, then $\{t_\beta^n(\alpha)\colon n \in \mathbf{Z}\}$ also fills S. The paper [52] includes some other results of this sort, but the following question remains open: is it possible to replace the Dehn twist t_α in this theorem by an arbitrary composition of Dehn twists about several disjoint circles, possibly at the cost of replacing the number 7 by some, may be undetermined, but *universal* number? The results of Fathi were preceded by a theorem of Long and Morton [140] in which only the finiteness of the number of exceptions (i.e., number of n such that $t_\alpha^n f$ is not pseudo-Anosov) was asserted, instead of the existence of a universal constant ($\leqslant 7$ by Theorem 7.1.D).

When one deals with individual mapping classes, one can usually reduce the problem to the case of irreducible mapping classes by the method of cutting outlined above. When one deals with the whole mapping class group or a general subgroup of it, usually there is no way to avoid reducible elements. But one can often restrict attention to following, in some sense simplest, type of reducible elements. Let us call a mapping class $f \in \mathrm{Mod}_S$ *pure* if it can be represented by a diffeomorphism $F: S \to S$ fixing (pointwise) some union C

of disjoint and pairwise non-isotopic nontrivial circles on S and such that F does not permute the components of $S \setminus C$ and induces on each component of the cut surface S_C a diffeomorphism isotopic either to a pseudo-Anosov or to the identity diffeomorphism. For example, every pseudo-Anosov element is pure, as is also any product of Dehn twists about several disjoint circles. In the latter case the induced diffeomorphism of S_C is isotopic to the identity; in all other cases F induces a diffeomorphism isotopic to a pseudo-Anosov diffeomorphism on at least one component of S_C.

THEOREM 7.1.E. *If $m \geqslant 3$, then $\Gamma_S(m)$ consists of pure elements.*

This theorem strengthens Corollary 7.1.B, and is a natural sharpening of it in the framework of Thurston's theory. For a proof, see [108, Corollary 1.8]. This theorem often allows the immediate elimination of all difficulties related to elements of finite order and to possible permutations of components (of the cut surface) by reducible diffeomorphisms. For example, if one considers a subgroup G of Mod_S, it is often possible to replace it by $G \cap \Gamma_S(m)$, $m \geqslant 3$.

Another important tool for dealing with reducible elements is the notion of essential reduction classes introduced, in a slightly different form, by Birman, Lubotzky and McCarthy [20]. For a pure element $f \in \mathrm{Mod}_S$ we say that a vertex α of $C(S)$ is an *essential reduction class* of f if $f(\alpha) = \alpha$ and $f(\beta) \neq \beta$ for all β such that α, β cannot be represented by disjoint circles. For a general element $f \in \mathrm{Mod}_S$, we say that a vertex α of $C(S)$ is an *essential reduction class* of f if it is an essential reduction class of some nontrivial (and than, as one may prove, of any nontrivial) power f^n ($n \neq 0$), which is pure. Note that f has pure powers in view of Theorem 7.1.E and the fact that $\Gamma_S(m)$ has finite index in Mod_S. It turns out that the set of all essential reduction classes of f is a simplex of $C(S)$, which we call the *canonical reduction system* of f and denote by $\sigma(f)$. If f is pure, it fixes all vertices of $\sigma(f)$. In general f leaves $\sigma(f)$ invariant. The crucial fact is that $\sigma(f)$ is non empty if f is a reducible element of infinite order ($\sigma(f)$ is empty for any element of finite order f, reducible or not). In particular, cutting along some realization of $\sigma(f)$ provides a canonical way to simplify a reducible element f. If f is pure, then this leads to a diffeomorphism (of the cut surface) which leaves all components invariant and is isotopic on any component either to the identity, or to a pseudo-Anosov diffeomorphism.

7.2. *Measuring against circles and its applications*

In this section we discuss one of the basic technical tools of Thurston's theory: measuring various geometric objects on a surface S against the isotopy classes of nontrivial circles, i.e., the vertices of $C(S)$. We will denote by $V(S)$ the set of all vertices of $C(S)$, i.e., the set of all isotopy classes of nontrivial circles. One of the fundamental ideas of Thurston was to consider measurements with respect to all vertices simultaneously. The natural place to record all such measurements is the set $\mathcal{R}(S)$ of all functions $V(S) \to \mathbf{R}_{\geqslant 0}$, endowed with the product topology. The mapping class group Mod_S acts on $V(S)$, and this action induces an action of Mod_S on $\mathcal{R}(S)$. The next important idea is to pass from the space $\mathcal{R}(S)$ to a projectivization of it, namely to the quotient $\mathcal{PR}(S)$ of $\mathcal{R}(S) \setminus \{0\}$ by the natural (diagonal) action of the multiplicative group $\mathbf{R}_{>0}$ on it. Clearly, the action of Mod_S on $\mathcal{R}(S)$ induces an action of Mod_S on $\mathcal{PR}(S)$.

Let us illustrate this approach by the example of the Teichmüller space T_S. A natural (but not the only possible) way to measure the points of T_S (i.e., the classes of hyperbolic metrics on S) with respect to the vertices $\alpha \in V(S)$ is to use the length functions $l_\alpha : T_S \to \mathbf{R}_{>0}$ from 5.2. The collection of all length functions defines an embedding $T_S \to \mathcal{R}(S)$. Clearly, this embedding is Mod_S-equivariant. Its image is contained, obviously, in $\mathcal{R}(S) \setminus \{0\}$. The composition $T_S \to \mathcal{PR}(S)$ of this embedding with the canonical projection $\mathcal{R}(S) \setminus \{0\} \to \mathcal{PR}(S)$ is still an embedding, which is also Mod_S-equivariant. It turns out that the embedding $T_S \to \mathcal{PR}(S)$ is a homeomorphism onto its image, this image has a compact closure in $\mathcal{PR}(S)$ and, moreover, this closure is homeomorphic to a closed disc (of the dimension equal to the dimension of T_S) under a homeomorphism taking the image of T_S into the interior of this disc. In this way, we achieve a natural compactification of T_S by a sphere called *Thurston's boundary* of Teichmüller space. The compactification itself is called *Thurston's compactification* of Teichmüller space. An advantage of this compactification is that the action of Mod_S extends to it by continuity (as is clear from the construction). Kerckhoff [121] proved that this desirable property does not hold for the classical compactifications of Teichmüller space by a sphere, the so-called *Teichmüller compactifications*. A Teichmüller compactification depends on a choice of a base point in T_S, and Kerckhoff's result holds for any choice of the base point. In fact, Kerckhoff proved that compactifications corresponding to different base points are often different, and this is closely related to the impossibility to extend the action of Mod_S to any of them by continuity.

Another example of the idea of measuring geometric objects against the isotopy classes of circles is provided by measuring the isotopy classes of circles themselves. For $\alpha, \beta \in V(S)$ let us define the *geometric intersection number* $\mathrm{i}(\alpha, \beta)$ as the minimal number of points in the intersections $A \cap B$, where A, B run over all circle representatives of the isotopy classes α, β respectively. For any $\alpha \in V(S)$ the function $I_\alpha : \gamma \mapsto \mathrm{i}(\gamma, \alpha)$ belongs to $\mathcal{R}(S)$. Moreover, it is easy to see that for any $\alpha \in V(S)$ there is some $\gamma \in V(S)$ such that $\mathrm{i}(\alpha, \gamma) \neq 0$. Hence, $I_\alpha \in \mathcal{R}(S) \setminus \{0\}$ for every $\alpha \in V(S)$. By assigning I_α to α, we get a map $V(S) \to \mathcal{R}(S) \setminus \{0\}$. It is easy to see that this map is injective. Moreover, its composition with the projection $\mathcal{R}(S) \setminus \{0\} \to \mathcal{PR}(S)$ is still injective (this is a little more difficult). The remarkable fact is that the image of $V(S)$ under this composition $V(S) \to \mathcal{PR}(S)$ is contained in Thurston's boundary of T_S and is dense in this boundary. In some sense, this image plays the role of rational points of Thurston's boundary.

It turns out that other points of Thurston's boundary also admit a geometric interpretation. The relevant geometric object is a *measured foliation with singularities*, or, more precisely, the *Whitehead equivalence class* of such a foliation on S. Usually one speaks simply about *measured foliations* on S. The set of all (Whitehead equivalence classes) of measured foliations on S is denoted by $\mathcal{MF}(S)$. We omit the definitions, referring to [53] and [222], and restrict ourselves to the following remarks. A measured foliation on S is a codimension 1 foliation on S (so the leaves have dimension 1). Since the Euler characteristic of S is almost never zero, S usually does not admit honest foliations, and we have to allow some singularities in them. The term *measured* means that our foliation is equipped with a so-called *transverse measure*, which allows us to measure locally (say, within a flow box of the foliation) the distance between the leaves. The transverse measure allows us to define the *transverse length* of curves and, in particular, of circles (if a curve is contained in a leaf, its transverse length is 0).

The notion of the transverse length of a circle allows us to deal with the measured foliations in a way similar to the way we dealt with hyperbolic metrics (or, rather, points of T_S) and circles (or, rather, vertices of $C(S)$) above. Namely, given a vertex $\gamma \in V(S)$ and a measured foliation μ, we define $l_\gamma(\mu)$ to be the infimum of transverse lengths with respect to μ of all circles in S representing the isotopy class γ. In contrast with the hyperbolic metrics (compare 5.2), this infimum is not always achieved (but it is achieved on a not necessarily embedded curve homotopic to circles representing γ). For any measured foliation μ, the function $L_\mu : \gamma \mapsto l_\gamma(\mu)$ belongs to $\mathcal{R}(S)$. It turns out that for every measured foliation μ there exists some $\gamma \in V(S)$ such that $l_\gamma(\mu) \neq 0$. Hence, $L_\mu \in \mathcal{R}(S) \setminus \{0\}$ for every $\mu \in \mathcal{MF}(S)$. By assigning L_μ to μ, we get a map $\mathcal{MF}(S) \to \mathcal{R}(S) \setminus \{0\}$. It is not very hard to see that this map is injective. In particular, we can equip $\mathcal{MF}(S)$ with the topology induced from $\mathcal{R}(S)$ by this map. The composition of this map with the projection $\mathcal{R}(S) \setminus \{0\} \to \mathcal{PR}(S)$ turns out to be not injective. The reason is very simple: the multiplication of transverse measures by positive constants defines a natural action of the multiplicative group $\mathbf{R}_{>0}$ on $\mathcal{MF}(S)$, and our map $\mathcal{MF}(S) \to \mathcal{R}(S) \setminus \{0\}$ is obviously $\mathbf{R}_{>0}$-equivariant. So, it is only natural to consider the quotient $\mathcal{PMF}(S)$ of $\mathcal{MF}(S)$ by this action of $\mathbf{R}_{>0}$. Clearly, our map $\mathcal{MF}(S) \to \mathcal{R}(S) \setminus \{0\}$ leads to a map $\mathcal{PMF}(S) \to \mathcal{PR}(S)$, which turns out to be injective. The image of this map is equal to Thurston's boundary of T_S. The interpretation of points of Thurston's boundary in terms of measured foliations plays a crucial role in the proof of the fact that Thurston's boundary is a sphere forming a ball together with the Teichmüller space.

Note that any isotopy class of a nontrivial circle naturally leads to an equivalence class of measured foliations, and the images in $\mathcal{R}(S)$ of an isotopy class of a circle and of the corresponding measured foliation are the same. So, a measured foliation turns out to be, in some sense, a generalized circle.

It is worth pointing out that Thurston's boundary can be constructed and its main properties can be established, using only the notion of a measured foliation without any reference to Teichmüller spaces, and for many applications it is sufficient to use only Thurston's boundary itself (together with the action of Mod_S on it).

7.3. *Action of the mapping classes on Thurston's compactification of Teichmüller space*

Let us denote by $\overline{T}_S$ Thurston's compactification of Teichmüller space, i.e., the union of T_S with its Thurston's boundary. In this section we describe the main features of the action of various classes of elements of Mod_S introduced in 7.1 on $\overline{T}_S$.

First of all, the elements of finite order can be characterized by the fact that they have a fixed point in the Teichmüller space T_S itself. The set of fixed points in T_S can be identified with some other Teichmüller space. A similar result probably holds for the set of fixed points in $\overline{T}_S$.

A pseudo-Anosov element f has exactly two fixed points in $\overline{T}_S$ and both of them belong to Thurston's boundary; there are no fixed points in T_S itself. One of the fixed points is in a natural sense *attracting*, and the other one is *repelling*. In the following theorem a denotes the attracting point and r denotes the repelling point.

THEOREM 7.3.A. *Let f be a pseudo-Anosov element. There exist two points a, r in Thurston's boundary of T_S, such that if U is any neighborhood of a in $\overline{T}_S$ and K is any compact set in $\overline{T}_S \setminus \{r\}$, then $f^n(K) \subset U$ for all sufficiently large n.*

So, all points in $\overline{T}_S \setminus \{r\}$ converge under the action of the iterations of f to a, uniformly on compact sets. This justifies the terms "attracting" and "repelling". Since both T_S and its boundary are obviously invariant, a similar result holds for Thurston's boundary instead of $\overline{T}_S$. One can find a somewhat weaker result (not asserting the uniformity of the convergence) in [53] (cf. [53, Exposé 12, Corollaire II.1]); Theorem 7.3.A itself easily follows, for example, from Theorem 7.3.B below.

The action of reducible elements is more complicated. If a reducible pure element f is not a composition of Dehn twists about several disjoint circles (so, if we cut S along a realization of $\sigma(f)$, then the induced diffeomorphism of the cut surface is isotopic to a pseudo-Anosov diffeomorphism on at least one component), then we have a picture similar to the above description of the action of a pseudo-Anosov element. We state the result only for the action on Thurston's boundary. Its set of fixed points consists of two components, one of which is in a natural sense *attracting*, while the other one is *repelling*. In the following theorem, A denotes the attracting set, and R denotes the repelling set.

THEOREM 7.3.B. *Let f be a pure element of* Mod_S *which is not a composition of Dehn twists about several disjoint circles. There exist two disjoint subsets A, R of Thurston's boundary of T_S, such that if U is any neighborhood of A in Thurston's boundary and K is any compact set in the complement of R, then $f^n(K) \subset U$ for all sufficiently large n.*

This result is proved in [108, Appendix], where also the components of the fixed point set are completely described. Theorem 7.3.A is a special case of Theorem 7.3.B and this description. Note that a weaker result, in which the attracting and the repelling sets are not disjoint in general (cf. [108, Theorem 3.5]), is sufficient for most applications. Another version of this weaker result was proved by McCarthy [159]; cf. [159, Uniform Convergence Lemma].

7.4. *Some applications*

We start with a famous result of Thurston about 3-manifolds, which in fact served as the motivation for Thurston's theory of surface diffeomorphisms. Recall that the mapping torus M_f of map $f : X \to X$ is the quotient space $X \times [0, 1]/(x, 0) \sim (f(x), 1)$. If X is a manifold and f is diffeomorphism, then M_f is also a manifold.

THEOREM 7.4.A. *Let $f : S \to S$ be a diffeomorphism of a closed surface S. The mapping torus M_f admits a Riemannian metric of constant negative curvature if and only if the isotopy class of f is pseudo-Anosov.*

A slightly more complicated result holds for surfaces with boundary as well. This theorem is the most difficult special case of Thurston's famous geometrization theorem for Haken manifolds. Of course, the proof involves much more than Thurston's theory of surface diffeomorphisms. The first account of this theorem is due to Sullivan [220]; the exposition in [220] is very demanding. Recently, new expositions of this theorem were published by McMullen [162] (see [162, Chapter 3]) and Otal [195]. McMullen's approach is close to the original ideas of Thurston (and refers to some still unpublished preprints of Thurston). The approach of Otal follows the one of Thurston in the general outline, but provides a completely new proof of one of the key steps (the so-called double limit theorem); his exposition is essentially self-contained. Hopefully, Thurston's own proof [223] will also appear soon. See also notes by Kapovich [120].

The next theorem, due to Kerckhoff [122,123], provides a solution of the so-called Nielsen realization problem.

THEOREM 7.4.B. *Let G be a finite subgroup of* Mod_S. *Then there is a finite subgroup* $G^{\sim}$ *of* $\mathrm{Diff}(S)$ *which maps isomorphically onto G under the natural map* $\mathrm{Diff}(S) \to \mathrm{Mod}_S$. *Moreover, one can choose* $G^{\sim}$ *to be a subgroup of the group of isometries of some metric of constant curvature on S with geodesic boundary.*

One says that $G^{\sim}$ *realizes* G. Note that this theorem generalizes Theorem 7.1.A from cyclic to arbitrary finite subgroups of Mod_S. The key ingredient of the proof is provided by the so-called *earthquake flows* on the Teichmüller space T_S. These flows are related to measured foliations in the same manner as the Fenchel–Nielsen flows from 5.2 are related to circles. In fact, earthquake flows are the limits of (suitably chosen and parameterized) Fenchel–Nielsen flows. The main properties of the earthquake flows used in the proof of Theorem 7.4.B are the convexity of the length functions from 5.2 along the flow lines of earthquake flows and the fact that every two points can be connected by the flow line of an earthquake flow. We refer to [123] for the details; see also [124]. Some further developments resulting from this solution of the Nielsen realization problem are presented in [125]. After Kerckhoff's work, Wolpert [234] and Tromba [228] (see [228, Section 6.4]), [229] found other proofs of Theorem 7.4.B, independent of Thurston's theory (but still based on convexity properties of some functions; in fact, the same length functions in Wolpert's proof, but along different paths in T_S).

The first applications of Thurston's theory to the algebraic properties of mapping class groups are due to McCarthy [158], who noticed that Thurston's theory provides a good description of normalizers and centralizers of pseudo-Anosov elements, and to Birman, Lubotzky and McCarthy [20], who proved, in particular, the following theorem.

THEOREM 7.4.C. *If a subgroup of* Mod_S *contains a solvable subgroup of finite index, then it contains an Abelian subgroup of finite index.*

Soon after [20], the author proved [95] the following more strong theorem.

THEOREM 7.4.D. *Let G be a subgroup of* Mod_S, *and let $m \geqslant 3$. Then either G contains a free (non-Abelian) group with two generators, or $G \cap \Gamma_S(m)$ is Abelian.*

Since $\Gamma_S(m)$ is of finite index in Mod_S and a solvable group cannot contain a free group with two generators, Theorem 7.4.D implies Theorem 7.4.C. The following corollary of Theorem 7.4.D, which is only slightly weaker than Theorem 7.4.D itself, was simultaneously and independently proved by McCarthy [159].

THEOREM 7.4.E. *Let G be a subgroup of* Mod_S. *Then either G contains a free group with two generators, or G contains an Abelian subgroup of finite index.*

A similar property, with "Abelian" replaced by "solvable", holds for arbitrary linear groups (i.e., subgroups of $\text{GL}_n(k)$ for a field k) according to a famous theorem of Tits [226]. In fact, the search for an analogue of Tits' theorem for subgroups of the mapping class groups led to Theorems 7.4.D and 7.4.E. Later on, the author [95,99] discovered that some other fundamental theorems about the linear groups have analogues for subgroups of mapping class groups. In order to state them, recall that a subgroup H of a group G is called *maximal* if for every subgroup K of G such that $H \subset K$, either $H = K$ or $K = G$. Recall also that the *Frattini subgroup* $F(G)$ of a group G is defined as the intersection of all maximal subgroups of G. An intuitive characterization of $F(G)$ is provided by the following elementary result: $F(G)$ is the set of all *nongenerating* elements of G, i.e., such elements $g \in G$ such that for every generating set X of G, the set $X \setminus \{g\}$ is also generating.

THEOREM 7.4.F. *Let G be a finitely generated subgroup of* Mod_S. *Then either G contains a maximal subgroup of infinite index, or G contains an Abelian subgroup of finite index.*

It is easy to see that if a group G contains an Abelian subgroup of finite index, then all maximal subgroups have finite index. While usually it is very easy to find a maximal subgroup of finite index (at least when there are nontrivial subgroups of finite index: take a subgroup of minimal index $\neq 1$), it is hard to prove the existence of maximal subgroups of infinite index. For linear groups, a theorem similar to Theorem 7.4.F (with "solvable" instead of "Abelian") was proved by Margulis and Soifer [151]; their result motivated Theorem 7.4.F.

THEOREM 7.4.G. *Let G be a finitely generated subgroup of* Mod_S. *Then the Frattini subgroup $F(G)$ is nilpotent.*

This is an analogue of a theorem of Platonov [203] to the effect that Frattini subgroups of finitely generated linear groups are nilpotent. It was proved by the author [99] after some partial results of Long [139].

The proofs of Theorems 7.4.C–7.4.G are based on the properties of the action of the elements of Mod_S on Thurston's boundary of the Teichmüller space T_S and, in particular, on the results discussed in 7.3. A crucial role is also played by the following result [97,99], in which we call a subgroup G of Mod_S *irreducible* if there is no nonempty simplex σ of $C(S)$ such that $g(\sigma) = \sigma$ for all $g \in G$.

THEOREM 7.4.H. *If G is an infinite irreducible subgroup of* Mod_S, *then G contains a pseudo-Anosov* (*in particular, an irreducible*) *element.*

Note that a similar statement for finite subgroups is false: according to Gilman [61], there are finite irreducible subgroups of Mod_S consisting entirely of reducible elements. In view of Theorems 7.4.B and 7.4.H, it is natural to classify the subgroups of Mod_S in a manner similar to Thurston's classification of elements of Mod_S. Namely, a subgroup of Mod_S can be *reducible* (i.e., not irreducible), finite, or infinite irreducible. (Note that these classes are not completely mutually exclusive: a subgroup can be finite and reducible.) The following theorem [108] provides a useful complement to Theorem 7.4.H and this classification.

THEOREM 7.4.I. *If a subgroup G of* Mod_S *contains a pseudo-Anosov element, then either G contains an infinite cyclic subgroup of finite index generated by a pseudo-Anosov element, or G contains a free group freely generated by two pseudo-Anosov elements.*

Theorems 7.4.C–7.4.G are proved using the same general outline. First, we replace our group G by $G \cap \Gamma_S(m)$ for some $m \geqslant 3$. If the resulting subgroup (still denoted by G) is reducible, then one may cut S along some realization of a nonempty simplex σ of $C(S)$ such that $g(\sigma) = \sigma$ for all $g \in G$. This reduces the problem to a similar problem on the cut surface (which has simpler components). If a subgroup G is irreducible and infinite, then Theorems 7.4.H and 7.4.I provide us with pseudo-Anosov elements in G, which play the key role in the rest of the proof. (Note that finite groups present no problems in these theorems.) For the details, see [108].

7.5. *Commuting elements and Dehn twists*

Thurston's theory allows us to formulate a very useful criterion for two elements of Mod_S to commute. We state it only for the pure elements, this case usually being sufficient for the applications.

THEOREM 7.5.A. *Let f, g be two pure elements of* Mod_S. *These elements commute (i.e., $fg = gf$) if and only if there is simplex σ of $C(S)$, its realization $C \subset S$ and representatives $F, G : S \to S$ of f, g respectively such that:*

(i) *F and G preserve (setwise) every component of C and every component of $S \setminus C$;*
(ii) *for every component R of the result S_C of cutting S along C, the diffeomorphisms $F_R, G_R : R \to R$ induced on R by F, G respectively are isotopic to some powers of the same pseudo-Anosov diffeomorphism of R or to the identity.*

The proof is based on the results of 7.3; see [108, Section 8.12]. As an application, we give a purely algebraic characterization of powers of Dehn twists inside subgroups of pure elements. Recall that the centralizer $C_G(g)$ of an element g of a group G is defined as the subgroup $\{f \in G\colon\ fg = gf\}$ and that the center $C(G)$ of a group G is defined as the subgroup $\{f \in G\colon\ fg = gf \text{ for all } g \in G\}$.

THEOREM 7.5.B. *Let G be a subgroup of* Mod_S *consisting entirely of pure elements (for example, let $G \subset \Gamma_S(m)$ for some $m \geqslant 3$). An element $g \in G$ is a nontrivial power of a Dehn twist if and only if the center $C(C_G(g))$ of the centralizer $C_G(g)$ of g is isomorphic to $\mathbf{Z}$ and is not equal to the centralizer $C_G(g)$ of g.*

In fact, $C(C_G(g))$ is isomorphic to $\mathbf{Z}$ only for nontrivial powers of Dehn twists and for pseudo-Anosov elements. The second condition serves to exclude pseudo-Anosov elements. This theorem is a version of a characterization of Dehn twists in [101].

Since powers of Dehn twists obviously are pure elements, one can apply Theorem 7.5.A to them and get the following result (admittedly, known before Thurston's theory).

THEOREM 7.5.C. *Let t_α, t_β be two Dehn twists and let m, n be two nonzero integers. Then $t_\alpha^n t_\beta^m = t_\beta^m t_\alpha^n$ if and only if α and β can be represented by disjoint circles.*

Theorem 7.5.A also allows us to easily find centers of the mapping class groups.

THEOREM 7.5.D. *If S is neither a sphere with $\leqslant 4$ holes, nor a torus with $\leqslant 2$ holes, nor a closed surface of genus 2, then the center $C(\mathrm{Mod}_S)$ is trivial (i.e., $C(\mathrm{Mod}_S) = 1$). If S is a closed surface of genus 2, then the center $C(\mathrm{Mod}_S)$ is isomorphic to $\mathbf{Z}/2\mathbf{Z}$ and is generated by the hyperelliptic involution from* 4.3.

The corresponding result for the cases excluded in this theorem is only slightly more complicated. Theorem 7.5.D is well known, but apparently there is no good reference for it.

COROLLARY 7.5.E. *If S is neither a sphere with $\leqslant 4$ holes, nor a torus with $\leqslant 2$ holes, nor a closed surface of genus 2, then the canonical action of $\mathrm{Mod}_S^\diamond$ on $C(S)$ is effective. If S is a closed surface of genus 2, the kernel of the canonical action of $\mathrm{Mod}_S^\diamond$ on $C(S)$ is equal to the subgroup of order 2 generated by the hyperelliptic involution from* 4.3.

This easily follows from Theorem 7.5.D, if one uses the fact that Dehn twists generate PMod_S (by Theorem 4.2.C).

8. Automorphisms of complexes of curves

8.1. *The theorem about automorphisms*

The action of $\mathrm{Mod}_S^\diamond$ on $C(S)$ induces a homomorphism $\mathrm{Mod}_S^\diamond \to \mathrm{Aut}(C(S))$. This homomorphism is injective if S is neither a sphere with $\leqslant 4$ holes, nor a torus with $\leqslant 2$ holes, nor a closed surface of genus 2; cf. Corollary 7.5.E. It turns out that this homomorphism is also surjective in these cases, and also for closed surfaces of genus 2.

THEOREM 8.1.A. *Suppose that S is not a sphere with $\leqslant 4$ holes and not a torus with $\leqslant 2$ holes. Then all automorphisms of $C(S)$ are induced by the elements of $\mathrm{Mod}_S^\diamond$. In particular, if, in addition, S is not a closed surface of genus 2, then $\mathrm{Aut}(C(S))$ is canonically isomorphic to $\mathrm{Mod}_S^\diamond$. If S is a closed surface of genus 2, then $\mathrm{Aut}(C(S))$ is canonically isomorphic to the quotient of the group $\mathrm{Mod}_S^\diamond$ by the subgroup of order 2 generated by the hyperelliptic involution.*

This theorem is due to the author [103] for surfaces S of genus $\geqslant 2$, and to Korkmaz [127] in the remaining cases. See also [110] for an expanded and updated version of [103]. Most of the proof (namely, the ideas discussed in 8.4) for the genus $\geqslant 2$ case works also in the genus 0 and 1 cases. The main difficulty in treating these low genus cases, overcome by Korkmaz, consisted in finding and proving an analogue of Lemma 8.2.A below.

If S is either a sphere with 4 holes, or a torus with $\leqslant 1$ holes, then $C(S)$ is an infinite set of vertices without any edges (i.e., $\dim C(S) = 0$) and, hence, $\mathrm{Aut}(C(S))$ is an infinite symmetric group. Obviously, diffeomorphisms of S cannot permute vertices of $C(S)$ arbitrarily, even in these cases, and thus the conclusion of the theorem is not true for such S. If S is a sphere with at most 3 holes, then $C(S)$ is empty, but $\mathrm{Mod}_S^\diamond$ is not trivial (it includes at least one orientation-reversing mapping class). Hence, the conclusion of the theorem is not true, by trivial reasons, for such S also. Luo [147] observed that $\mathrm{Aut}(C(S))$ is not equal to Mod_S if S is a torus with 2 holes. The reason is very simple: If $S_{1,2}$ is a torus with 2 holes and $S_{0,5}$ is a sphere with 5 holes, then $C(S_{1,2})$ is isomorphic to $C(S_{0,5})$, but $\mathrm{Mod}_{S_{1,2}}$ is not isomorphic to $\mathrm{Mod}_{S_{0,5}}$; in fact, $\mathrm{Mod}_{S_{0,5}}$ acts transitively on the set of vertices of $C(S_{0,5})$ and the action of $\mathrm{Mod}_{S_{1,2}}$ on the set of vertices of $C(S_{1,2})$ has two orbits, corresponding to separating and nonseparating circles (cf. 8.6 for a similar phenomenon). Luo [147] also suggested another approach to Theorem 8.1.A, still based on the ideas of [103] (outlined below) and also on a multiplicative structure on the set of vertices of $C(S)$ introduced in [145].

Theorem 8.1.A is similar to, and was motivated by, a well known theorem of Tits [227] to the effect that all automorphisms of a Tits building (with some exceptions, as in our case) stem from automorphisms of the corresponding algebraic group; see [227, Introduction, Problem (B)]. The theorem of Tits, in turn, has its roots in the *Fundamental Theorem of Projective Geometry*, asserting that maps of a projective space to itself which map lines to lines, planes to planes etc., are, in fact, (projectively) linear.

In 8.2–8.4 we will outline the main ideas of the proof of Theorem 8.1.A. After this, we will discuss applications of this result to mapping class groups and Teichmüller spaces.

8.2. *Intersection number 1 property*

We start by introducing some terminology. Two isotopy classes α, β of non-trivial circles on S (i.e., two vertices of $C(S)$) are said to have *geometric intersection number* 1 (respectively, 0) if they can represented by two circles intersecting transversely in exactly one point (respectively, by two disjoint circles). If this is the case, we write $\mathrm{i}(\alpha, \beta) = 1$ (respectively, $\mathrm{i}(\alpha, \beta) = 0$). (These definitions form a special case of a general notion of geometric intersection number from 7.2.) Clearly, two isotopy classes have geometric intersection number 0 if and only if they are connected by an edge in $C(S)$. Obviously, these intersection numbers are preserved by the action of $\mathrm{Mod}_S^\diamond$. The starting point of the proof of Theorem 8.1.A (at least for surfaces of genus $\geqslant 1$) is the fact that the property of having geometric intersection number 1 is preserved by all automorphisms of $C(S)$ (which a priori preserve only the property of having geometric intersection number 0). The key part of the proof of this fact is contained in the following lemma, at least when the genus is $\geqslant 2$.

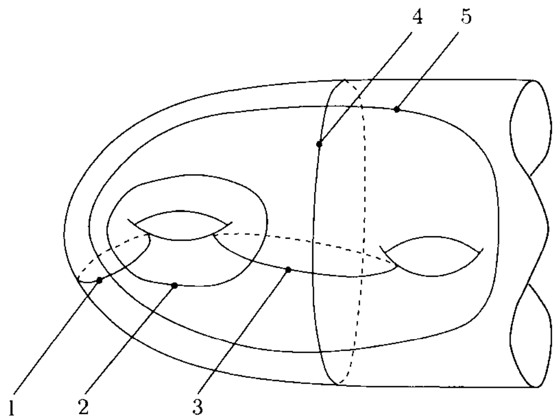


Fig. 14.

LEMMA 8.2.A. *Suppose that the genus of S is at least 2. Let α_1, α_2 be isotopy classes of two nontrivial circles on S. Then the geometric intersection number* $\mathrm{i}(\alpha_1, \alpha_2) = 1$ *if and only if there exist isotopy classes α_3, α_4, α_5 of nontrivial circles having the following two properties*:

(i) $\mathrm{i}(\alpha_i, \alpha_j) = 0$ *if and only if the i-th and j-th circles in Figure* 14 *are disjoint*;

(ii) *if α_4 is the isotopy class of a circle C_4, then C_4 divides S into two parts, one of which is a torus with one hole containing some representatives of the isotopy classes α_1, α_2.*

PROOF. Figure 14 provides a proof of the "only if" part. In view of the property (ii), the proof of the "if" part can be reduced to an examination of a torus with one hole, which is not difficult. We omit the details. □

COROLLARY 8.2.B. *Suppose that the genus of S is at least 2. Every automorphism of $C(S)$ maps pairs of vertices having geometric intersection number 1 into pairs of vertices also having this property.*

PROOF. It is sufficient to show that the properties (i) and (ii) of Lemma 8.2.A are preserved by every automorphism of $C(S)$. As we noted before the lemma, two vertices have geometric intersection number 0 if and only if they are connected by an edge in $C(S)$. It follows that the property (i) is preserved by all automorphisms. In order to prove that the property (ii) is also preserved, we will show how to express it in terms of the structure of $C(S)$ only.

Let us consider a vertex $\alpha = \langle C \rangle$ of $C(S)$. Let L_α be the link of α in $C(S)$, i.e., the set of all vertices of $C(S)$ connected by an edge with α (with the induced structure of a simplicial complex). Consider the graph L_α^*, having the same vertices as the link L_α and having as edges exactly those pairs of vertices which are *not* edges of L_α. It is easy to see that the connected components of L_α^* correspond exactly to the connected components of the result S_C of cutting S along C (at least if none of them is a sphere with 3 holes).

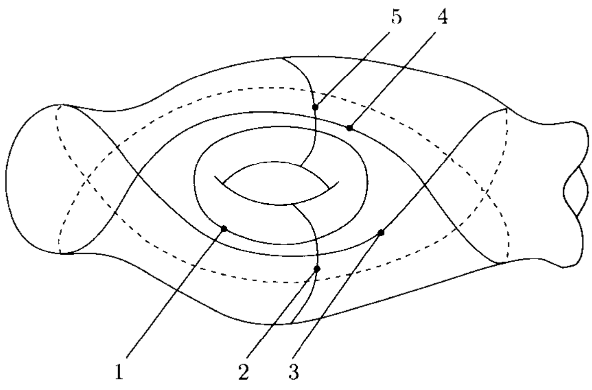


Fig. 15.

After detecting the components of S_C with the help of L_α^* (when one of the components is a sphere with 3 holes, we can detect only the other, but this turns out to be sufficient), we can return to the link L_α itself and try use Corollary 3.3.B, or the table preceding it, to recognize the topological type of these components.

Let us apply these remarks to α_5. The part of S *not* containing representatives of α_1, α_2 should have genus one less than the genus of S and one more boundary component. If the genus of S is $\geqslant 3$, these properties can be recognized with the help of Corollary 3.3.B. The case of genus 2 is only a little more difficult; one should use not only this corollary, but also the table preceding it. It follows that the property (ii) can be expressed in terms of the structure of $C(S)$ alone and, hence, is preserved by all automorphisms of $C(S)$. This completes the proof. □

Now, we turn to the genus 1 case. The following analogue of Lemma 8.2.A was proved by Korkmaz [127].

LEMMA 8.2.C. *Suppose that S is a torus with at least two holes. Let α_1, α_2 be isotopy classes of two nontrivial circles on S. Then the geometric intersection number* $\mathrm{i}(\alpha_1, \alpha_2) = 1$ *if and only if there exist isotopy classes α_3, α_4, α_5 having the following two properties*:

(i) $\mathrm{i}(\alpha_i, \alpha_j) = 0$ *if and only if the i-th and j-th circles in Figure* 15 *are disjoint*;

(ii) *if α_i is the isotopy class of a circle C_i, then the circles C_1, C_2 and C_3 are non-separating, and both C_4 and C_5 divide S into a torus with one hole and a sphere with $b + 1$ holes, where b is the number of boundary components of S.*

PROOF. Similarly to the proof of Lemma 8.2, the proof of the"only if" part is provided by the Figure 15. The proof of the "if" part is a little more difficult than the proof of the "if" part of Lemma 8.2, but we again omit the details. □

COROLLARY 8.2.D. *Suppose that S is a torus with at least two holes. Every automorphism of $C(S)$ maps pairs of vertices having geometric intersection number* 1 *into pairs of vertices also having this property.*

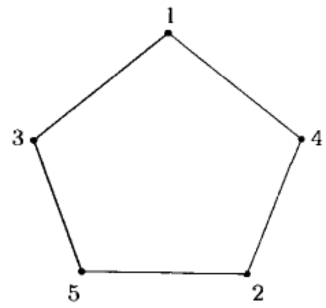


Fig. 16.

The proof is more difficult than that of Corollary 8.2.B, partly because the table preceding Corollary 3.3.B is not sufficient to distinguish between the surfaces of genus 0 and 1. We omit fairly complicated details.

Surprisingly, the vertices $\alpha_1, \ldots, \alpha_5$ of both Lemmas 8.2.A and 8.2.C with the edges connecting them form a pentagon, presented in Figure 16, and a similar pentagon plays a role in the case of genus 0.

8.3. *The case of a sphere with holes*

At first sight, none of the above ideas can be applied when S is a sphere with holes: all circles on a sphere with holes are separating and, hence, two isotopy classes never have geometric intersection number 1. It turns out that instead of non-separating circles, one can use circles bounding a disc with two holes in S. Then the role of pairs of isotopy classes having geometric intersection number 1 is taken over by the pairs of isotopy classes of such circles having the simplest possible non-trivial intersection, namely, intersecting as in Figure 17, where the shadowed discs represent holes. Such pairs of isotopy classes can be characterized in a manner similar to Lemmas 8.2.A and 8.2.C. A pentagon as in Figure 16 appears again, but its five vertices are not sufficient now. We refer to [127] for the details.

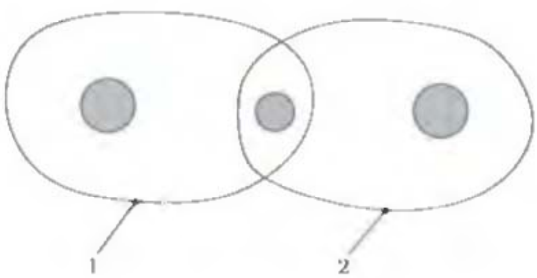


Fig. 17.

Note that the circles bounding discs with two holes and pairs illustrated in Figure 17 are exactly what is needed to make the machinery of the next section work (cf. the discussion of codings in 8.4).

8.4. *Complexes of curves and ideal triangulations*

Now we describe the ideas involved in the remaining part of the proof of Theorem 8.1.A. These ideas work (almost) equally well for surfaces of any genus. We restrict ourselves to the genus $\geqslant 2$ case.

We need a minor modification $A(S)$ of the ideal triangulations of the Teichmüller spaces from 5.5. The vertices of the complex $A(S)$ are the isotopy classes $\langle I \rangle$ of *properly embedded non-trivial* (cf. 2.3) arcs I in S. The isotopies are not required to be fixed at the ends, but are required to run in the class of properly embedded arcs. A set of vertices $\{\gamma_0, \ldots, \gamma_n\}$ forms a simplex of $A(S)$ if and only if the isotopy classes $\gamma_0, \ldots, \gamma_n$ can be represented by pairwise disjoint arcs $I_0, \ldots, I_n$. The extended mapping class group $\mathrm{Mod}_S^\diamond$ acts on $A(S)$ in an obvious way.

We can collapse each boundary component of S into a point (different for different boundary components) and then consider these points as punctures. Let R be the resulting surface with punctures. There is a natural map $S \to R$, establishing a diffeomorphism between the complements of the boundary and of the punctures. Clearly, this map induces an isomorphism between $A(S)$ and the complex $A(R)$ from 5.5. Using this isomorphism and Corollary 5.5.B, we immediately get the following result.

LEMMA 8.4.A. *Every automorphism of $A(S)$ is determined by its action on any single top dimensional simplex. In particular, if an automorphism of $A(S)$ agrees with the action of some element of $\mathrm{Mod}_S^\diamond$ on some simplex of codimension 0, then this automorphism agrees with this element of $\mathrm{Mod}_S^\diamond$ on the whole complex $A(S)$.*

In view of the last lemma, the automorphisms of $A(S)$ seem to be much more accessible than the automorphisms of $C(S)$. The main idea of the remaining part of the proof of Theorem 8.1.A is to reduce the study of automorphisms of $C(S)$ to a study of automorphisms of $A(S)$. The main tool is a coding of the vertices of $A(S)$ in terms of the vertices of $C(S)$ and some additional data.

Let us describe this coding. The vertices of $A(S)$ are naturally divided into three types, and a vertex is coded by its type and one or two vertices of $C(S)$, depending on its type. The types of vertices are the following: (a) the isotopy classes of arcs I connecting two different components D_1, D_2 of ∂S; (b) the isotopy classes of arcs I connecting a component D of ∂S with itself such that the image of this arc in R (this image is a circle passing through a puncture) does not bound a disc with one puncture; (c) the isotopy classes of arcs I connecting a component D of ∂S with itself such that the image of this arc in R does bound a disc with one puncture. In addition to the type, a vertex of type (a) is coded by $\langle C \rangle$, where C is the boundary of some regular neighborhood of $D_1 \cup I \cup D_2$ in S, a vertex of type (b) is coded by the pair $\langle C_1 \rangle$, $\langle C_2 \rangle$, where C_1, C_2 are two components of a regular neighborhood of $D \cup I$, and a vertex of type (c) is coded by $\langle C \rangle$, where C is the non-trivial component of a regular neighborhood of $D \cup I$ (one of the components is trivial in this case).

Now, we need to prove that every automorphism of $C(S)$ maps vertices coding vertices of $A(S)$ of type (a) into similar vertices, behaves similarly with respect to types (b) and (c) and also maps codings of pairs of vertices connected by an edge in $A(S)$ into similar codings. Figure 17 may serve as an illustration of codings of a pair of vertices connected by an edge. This is the most technically difficult part of the proof, which splits into many cases according to the types of the vertices involved. After this is done, we can assign an automorphism of $A(S)$ to any automorphism of $C(S)$. Moreover, one can prove that on an explicitly given codimension 0 simplex of $A(S)$ every automorphism of $A(S)$ induced by an automorphism of $C(S)$ agrees with an element of $\mathrm{Mod}_S^\diamond$, at least if the number of boundary components is $\geqslant 2$. In this case Lemma 8.4.A implies that the induced automorphism of $A(S)$ is equal to some element of $\mathrm{Mod}_S^\diamond$. And if the induced automorphism of $A(S)$ is equal to some element of $\mathrm{Mod}_S^\diamond$, then the original automorphism of $C(S)$ is equal to some element of $\mathrm{Mod}_S^\diamond$, as it is easy to see.

This proves Theorem 8.1.A for surfaces with $\geqslant 2$ boundary components. The case of surfaces with $\leqslant 1$ boundary components can be reduced to the case of surfaces with $\geqslant 2$ boundary components as follows. Note first that every automorphism of $C(S)$ takes the isotopy classes of non-separating circles into isotopy classes of non-separating circles (for example, because C is non separating if and only if there exist a vertex β such that $i(\alpha, \beta) = 1$, where $\alpha = \langle C \rangle$). Since all non-separating circles are in the same orbit of the group of diffeomorphisms of S, we can assume that our automorphism of $C(S)$ fixes some vertex $\alpha = \langle C \rangle$, where C is non-separating. Such an automorphism induces an automorphism of the link L_α of α and, hence, of the complex $C(S_C)$, where S_C is the result of cutting S along C (compare with the proof of Corollary 8.2.B). Since S_C has $\geqslant 2$ boundary components, this automorphism of $C(S_C)$ is equal to some element of $\mathrm{Mod}_{S_C}^\diamond$. Considering different non-separating circles C (in fact, all of them), one can deduce that the original automorphism of $C(S)$ is equal to some element of $\mathrm{Mod}_S^\diamond$. (If the genus of S is 2, an additional difficulty is caused by the fact that the genus of S_C will be < 2. But our automorphisms of $C(S_C)$ are induced by automorphisms of $C(S)$ and, hence, retain all the nice properties of the latter, such as the preservation of the intersection number 1.)

This completes our outline of the proof of Theorem 8.1.A.

8.5. *An application to subgroups of finite index*

As the first application of Theorem 8.1.A, we give a complete description of isomorphisms between subgroups of finite index in $\mathrm{Mod}_S^\diamond$. This result will find its own applications in 9.2; cf. Corollary 9.2.B.

THEOREM 8.5.A. *Suppose that S is not a sphere with $\leqslant 4$ holes and not a torus with $\leqslant 2$ holes. Let Γ_1 and Γ_2 be two subgroups of finite index of* $\mathrm{Mod}_S^\diamond$. *If S, in addition, is not a closed surface of genus 2, then all isomorphisms $\Gamma_1 \to \Gamma_2$ have the form $x \mapsto gxg^{-1}$ for some $g \in \mathrm{Mod}_S^\diamond$. If S is a closed surface of genus 2, then all isomorphisms $\Gamma_1 \to \Gamma_2$ have the form $x \mapsto gxg^{-1}\chi(x)$ for some $g \in \mathrm{Mod}_S^\diamond$ and some homomorphism χ from Γ_1*

to the center of $\mathrm{Mod}_S^\diamond$, *i.e., the subgroup of* $\mathrm{Mod}_S^\diamond$ *of order* 2 *generated by the hyperelliptic involution* (*see Theorem* 7.5.D); χ *can be non-trivial only if* Γ_2 *contains the hyperelliptic involution.*

PROOF. First, note that for some natural number $N \neq 0$ the N-th powers t_α^N of all Dehn twists t_α are contained in Γ_1, because Γ_1 is of finite index. Next, Theorem 7.5.B easily implies that any isomorphism $\varphi : \Gamma_1 \to \Gamma_2$ maps sufficiently high powers of Dehn twists into powers of Dehn twists. Taking into account the fact that two non-trivial powers of Dehn twists commute if and only if the corresponding isotopy classes of circles have geometric intersection number 0 (by Theorem 7.5.C), we see that every such isomorphism φ induces an automorphism $C(S) \to C(S)$. By Theorem 8.1.A, this automorphism of $C(S)$ is given by some element $g \in \mathrm{Mod}_S^\diamond$. This means that for some sufficiently large N, we have

$$\varphi\left(t_\alpha^N\right) = t_{g(\alpha)}^{M_\alpha}$$

for some $M_\alpha \neq 0$, for all vertices α of $C(S)$. The potential dependence of M_α on α is irrelevant in what follows, and we write simply M for M_α.

Now, let $f \in \Gamma_1$. Then, for any vertex α,

$$\varphi\left(f t_\alpha^N f^{-1}\right) = \varphi\left(t_{f(\alpha)}^N\right) = t_{g(f(\alpha))}^M.$$

On the other hand,

$$\varphi\left(f t_\alpha^N f^{-1}\right) = \varphi(f)\varphi\left(t_\alpha^N\right)\varphi(f)^{-1} = \varphi(f) t_{g(\alpha)}^M \varphi(f)^{-1} = t_{\varphi(f)(g(\alpha))}^M.$$

Comparing the results of these two computations, we conclude (note that two nontrivial powers t_α^n, t_β^m of right Dehn twists are equal if and only if $\alpha = \beta$ and $m = n$) that $\varphi(f)(g(\alpha)) = g(f(\alpha))$ for all α and (after putting $\alpha = g^{-1}(\beta)$) that $\varphi(f)(\beta) = gfg^{-1}(\beta)$ for all vertices β of $C(S)$. If S is not a closed surface of genus 2, this implies that $\varphi(f) = gfg^{-1}$, i.e., φ has the required form; see Corollary 7.5.E. If S is a closed surface of genus 2, then $\varphi(f)$ can differ from gfg^{-1} by the hyperelliptic involution. It is easy to see that this difference can be described by some homomorphism χ as stated. □

COROLLARY 8.5.B. *Suppose that S is not a sphere with* $\leqslant 4$ *holes and not a torus with* $\leqslant 2$ *holes. Let Γ be a subgroup of finite index of* $\mathrm{Mod}_S^\diamond$. *Then the outer automorphism group* $\mathrm{Out}(\Gamma)$ *is finite.*

Theorem 8.5.A was proved first by the author [95] for the case $\Gamma_1 = \Gamma_2 = \mathrm{Mod}_S^\diamond$ or Mod_S and for closed surfaces only. Another, but closely related, proof in this case was provided by McCarthy [160], who also noticed additional automorphisms resulting from the hyperelliptic involution in the genus 2 case (which were overlooked in [95]). An alternative approach was suggested by Tchangang [221]. Shortly after [95] the author extended these results to surfaces with boundary and to some natural subgroups of $\mathrm{Mod}_S^\diamond$ like $\Gamma_1 = \Gamma_2 = \mathrm{PMod}_S$; see [101]. A crucial element of these early approaches, namely the fact that any isomorphism φ takes powers of Dehn twists to powers of Dehn twists, also plays a crucial role in the above proof of Theorem 8.5.A.

Recently, McCarthy and the author extended the results of [101,160] to some injective homomorphisms between mapping class groups. Let us measure the *size* of a surface S by the number $3g-3+b$, where g is the genus and b is the number of boundary components of S (this number is equal to the maximal number of pairwise nonisotopic disjoint circles on S). The main result of [111,112] asserts that if the sizes of two surfaces S, S' differ by at most 1 and they have the genus $\geqslant 1$, then injective homomorphisms $\mathrm{Mod}_S \to \mathrm{Mod}_{S'}$ are, in fact isomorphisms and S is diffeomorphic to S' with, possibly, a few exceptions.

8.6. *An application to Teichmüller spaces*

In this section we consider the Teichmüller spaces of *punctured* surfaces (cf. 5.1).

The Teichmüller space T_R of a punctured surface R carries a natural structure of a metric space, given by its *Teichmüller metric* (we note in passing that Teichmüller spaces have a couple of other natural metrics). This structure of a metric space is not derived from a Riemannian metric (although it can be derived from a Finslerian metric), but nevertheless has some nice geometric properties. In particular, T_R is a complete metric space, every two points can be connected by a geodesic (i.e., a locally shortest path), and, moreover, geodesics have a nice description in terms of the (conformal) geometry of surfaces. For an introduction to this theory, see Abikoff's book [1]. The fact that the metric is naturally defined implies that Mod_R° acts on T_R by isometries. A remarkable result proved by Royden [208] for closed surfaces (and isometries preserving the natural complex structure of T_S), and then extended by Earl and Kra [47] to the surfaces with boundary (and general isometries), asserts that there are no other isometries. More precisely, we have the following theorem.

THEOREM 8.6.A. *Suppose that R is not a sphere with $\leqslant 4$ punctures and not a torus with $\leqslant 1$ punctures. Then any isometry of T_R is induced by some element of* $\mathrm{Mod}_R^\diamond$.

Note that the conclusion of this theorem is not true for a sphere with 4 punctures and a torus with 0 or 1 punctures, because for these surfaces the Teichmüller space is isometric to the hyperbolic plane, which has a continuous group of isometries. For a sphere with $\leqslant 3$ punctures the conclusion is vacuous, because the Teichmüller space consists of just one point. For a torus with 2 punctures the conclusion of the theorem is not true, because if $S_{1,2}$ is a torus with 2 punctures and $S_{0,5}$ is a sphere with 5 punctures, then $T_{S_{1,2}}$ is isometric to $T_{S_{0,5}}$, but $\mathrm{Mod}_{S_{1,2}}$ is not isomorphic to $\mathrm{Mod}_{S_{0,5}}$ (cf. 8.1). See [47].

Royden's proof (and its extension by Earl and Kra) is analytic and is based on a detailed investigation of the non-smoothness properties of the Finslerian metric underlying the Teichmüller metric. It is essentially local in nature. A detailed exposition of this proof is presented in Gardiner's book [58]; cf. [58, Chapter 9]. Our Theorem 8.1.A can be used to prove this theorem in a completely different manner (in all cases except that of a torus with two punctures). This new proof, outlined in [103], reveals a deep analogy between Royden's theorem and Mostow's rigidity theorems [184,185]. In fact, Theorem 8.1.A plays a role in this proof similar to the role played by the theorem of Tits mentioned in 8.1 in Mostow's proof [185]. In addition to Theorem 8.1.A this proof of Theorem 8.6.A, naturally, relies heavily on the theory of the Teichmüller spaces and, in particular, on the results of Kerckhoff [121] and Masur [152–154]. Unfortunately, this theory falls outside the scope of the present paper.

9. Mapping class groups and arithmetic groups

9.1. *Arithmetic groups*

The analogy between arithmetic groups and mapping class groups was suggested first by Harvey [86], and had guided much of research about the mapping class groups since then. Let us give a down-to-earth version of the definition of arithmetic groups.

The basic example of an arithmetic group is $\mathrm{SL}_n(\mathbf{Z})$, the group of all integral $n \times n$ matrices with the determinant 1 (note that the inverse of an integral matrix with the determinant 1 is automatically an integral matrix). All other arithmetic groups can be, in some sense, constructed from this one. Suppose that G is a *semisimple* **Q**-*algebraic* subgroup of $\mathrm{SL}_n(\mathbf{R})$. This means that G is simultaneously a semisimple Lie subgroup of $\mathrm{SL}_n(\mathbf{R})$ and a subset of the space of all real $n \times n$ matrices which can be described by polynomial equations for the matrix entries with *rational* coefficients. For any such G the intersection $G_{\mathbf{Z}}$ of G with the set of integral matrices is, by definition, an arithmetic group. Next, let G_0 be the connected component of the identity of G and $\varphi : G_0 \to H$ be a homomorphism from G_0 onto a connected semisimple Lie group H with compact kernel (to be honest, we should also assume that H has trivial center and no compact factors, but the reader may ignore this remark). Then $\Gamma = \varphi(G_{\mathbf{Z}} \cap G_0)$ is, by definition, an arithmetic group, as is also any subgroup Γ' of H *commensurable* with it, i.e., any subgroup Γ' such that the intersection $\Gamma \cap \Gamma'$ has finite index both in Γ and Γ'. More precisely, we say that any such Γ' is an *arithmetic subgroup of* H. We also call *arithmetic* any group isomorphic to any such Γ'.

So, any arithmetic group can be obtained from $\mathrm{SL}_n(\mathbf{Z})$ in three steps: taking an intersection with a semisimple **Q**-algebraic real Lie group, then taking the image under a surjective homomorphism with compact kernel (this step is the most difficult to control), and, finally, taking any subgroup commensurable with this image. In particular, any subgroup of finite index of an arithmetic group is also an arithmetic group.

In addition to $\mathrm{SL}_n(\mathbf{Z})$, a typical example of an arithmetic group is the symplectic group $\mathrm{Sp}_{2g}(\mathbf{Z})$, which shows up in the topology of surfaces as the group of all automorphisms of the first homology group $H_1(S_g)$ of a closed orientable surface S_g of genus g preserving the intersection pairing. In particular, we have a natural homomorphism $\mathrm{Mod}_{S_g} \to \mathrm{Sp}_{2g}(\mathbf{Z})$, which is well known to be surjective.

In [86], Harvey asked if the mapping class groups are arithmetic. They turned out to not be, as was proved by the author [95,100]. On the other hand, the years after [86] profoundly demonstrated that there is indeed a deep analogy between mapping class groups and arithmetic groups, especially in their cohomology properties (the main results of Section 6 are, in fact, motivated by analogous results about arithmetic groups). Although the first proof of the non-arithmeticity of the mapping class groups was quite short (cf. 9.3), it did not exhaust the problem, and the question of *why* the mapping class groups are not arithmetic led to further results. In the next section (cf. 9.2) we will describe what seems to be the best answer to this question.

9.2. *Abstract commensurators*

The notion of the abstract commensurator of a group has its roots in the theory of arithmetic groups. The author learned it from [190]; see also [5, Appendix B], where it is attributed to Serre and Neumann. It allows us to express one of the most striking differences between arithmetic and non-arithmetic groups; cf. Theorem 9.2.A below and the remarks after it.

Let Γ be a group. Let us consider all possible isomorphisms $\varphi : \Gamma_1 \to \Gamma_2$ between subgroups Γ_1, Γ_2 of finite index of Γ. Let us identify two such isomorphisms φ, φ' defined on Γ_1, Γ_1' respectively, if they agree on a subgroup of finite index in the intersection $\Gamma_1 \cap \Gamma_1'$. We can compose them in an obvious manner; the composition $\varphi \circ \varphi'$ of $\varphi : \Gamma_1 \to \Gamma_2$ with $\varphi' : \Gamma_1' \to \Gamma_2'$ is defined on $\varphi^{-1}(\Gamma_2 \cap \Gamma_1')$. Under this composition operation, the classes of such isomorphisms form a group, which is called the *abstract commensurator* of Γ and is denoted by $\mathrm{Comm}(\Gamma)$. There is a natural map $i : \Gamma \to \mathrm{Comm}(\Gamma)$, sending an element γ of Γ to the (class of the) inner automorphism $g \mapsto \gamma g \gamma^{-1}$. This map is injective if the centralizers of subgroups of finite index in Γ are trivial, as is the case for arithmetic groups and for mapping class groups (see Theorem 7.5.D for the latter).

As a nice example, let us mention that $\mathrm{Comm}(\mathbf{Z}^n) = \mathrm{SL}_n(\mathbf{Q})$.

THEOREM 9.2.A. *If Γ is an arithmetic group, then $i(\Gamma)$ is of infinite index in* $\mathrm{Comm}(\Gamma)$.

PROOF. We only outline the main idea. Let us consider $\mathrm{SL}_n(\mathbf{Z})$ and show that any element of $\mathrm{SL}_n(\mathbf{Q})$ naturally defines an element of $\mathrm{Comm}(\mathrm{SL}_n(\mathbf{Z}))$.

Let $g \in \mathrm{SL}_n(\mathbf{Q})$, and let m be the least common multiple of all denominators of the entries of the matrices g, g^{-1}. Let $\Gamma_1 = \{\gamma \in \mathrm{SL}_n(\mathbf{Z}) \mid \gamma \equiv I \bmod m^2\}$, where I is the identity matrix. Clearly, Γ_1 is of finite index in $\mathrm{SL}_n(\mathbf{Z})$. If $\gamma \in \Gamma_1$, then $\gamma = I + m^2 h$, where h is an integral matrix, and hence $g\gamma g^{-1} = I + m^2 g h g^{-1}$ is also an integral matrix. Thus $g\Gamma_1 g^{-1} \subset \mathrm{SL}_n(\mathbf{Z})$. Let us show that $\Gamma_2 = g\Gamma_1 g^{-1}$ is a subgroup of finite index in $\mathrm{SL}_n(\mathbf{Z})$. Consider the subgroup $\Gamma' = \{\gamma \in \mathrm{SL}_n(\mathbf{Z}) \mid \gamma \equiv I \bmod m^4\}$. If $\gamma \in \Gamma'$, then $\gamma = I + m^4 h$, where h is an integral matrix, and hence $g^{-1}\gamma g = I + m^2(m^2 g^{-1} h g) \in \Gamma_1$, because $m^2 g^{-1} h g$ is an integral matrix. It follows that $g^{-1}\Gamma' g \subset \Gamma_1$, and hence $\Gamma' \subset g\Gamma_1 g^{-1}$. Since Γ' obviously has finite index in $\mathrm{SL}_n(\mathbf{Z})$, we conclude that $\Gamma_2 = g\Gamma_1 g^{-1} \supset \Gamma'$ is of finite index.

Now, it is clear that the inner automorphism $x \mapsto gxg^{-1}$ induces an isomorphism $\Gamma_1 \to \Gamma_2$ between two subgroups of finite index of $\mathrm{SL}_n(\mathbf{Z})$, and hence an element of $\mathrm{Comm}(\mathrm{SL}_n(\mathbf{Z}))$. This leads to a homomorphism $\mathrm{SL}_n(\mathbf{Q}) \to \mathrm{Comm}(\mathrm{SL}_n(\mathbf{Z}))$. Clearly, its image contains $i(\mathrm{SL}_n(\mathbf{Z}))$ as a subgroup of infinite index. This proves the theorem for $\mathrm{SL}_n(\mathbf{Z})$.

The same argument works for the groups $G_{\mathbf{Z}}$, where G is a semisimple $\mathbf{Q}$-algebraic group, simply because $G_{\mathbf{Z}}$, like $\mathrm{SL}_n(\mathbf{Z})$, has a corresponding group of rational matrices associated to it. After this, it is not hard to extend the result to the groups of the form $\varphi(G_{\mathbf{Z}} \cap G_0)$ (cf. 9.1) and then to any group commensurable with such a group. We refer to Zimmer's book [246] for the details; cf. [246, Section 6.2]. □

A converse to this theorem is also true: if Γ is a *lattice* in a semisimple Lie group G (i.e., if G/Γ has finite invariant volume) and $i(\Gamma)$ is of infinite index in $\mathrm{Comm}(\Gamma)$, then Γ is arithmetic. This result is an immediate corollary of an arithmeticity theorem of Margulis (cf. [246, Chapter 6]) and Mostow's rigidity theorem [184,185]; cf. [190, Section 1]. This converse is much more deep and difficult than Theorem 9.2.A, and we do not need it for our applications.

COROLLARY 9.2.B. *Suppose that S is not a sphere with $\leqslant 4$ holes and not a torus with $\leqslant 2$ holes. Then the mapping class groups Mod_S, $\mathrm{Mod}_S^\diamond$, and all subgroups of finite index in them, are not arithmetic.*

PROOF. It follows from Theorem 8.5.A that $\mathrm{Comm}(\mathrm{Mod}_S) = \mathrm{Comm}(\mathrm{Mod}_S^\diamond) = i(\mathrm{Mod}_S^\diamond)$, and hence Mod_S, $\mathrm{Mod}_S^\diamond$ cannot be arithmetic by Theorem 9.2.A.

If Γ is a subgroup of finite index in $\mathrm{Mod}_S^\diamond$, then again $\mathrm{Comm}(\Gamma) = i(\mathrm{Mod}_S^\diamond)$ by Theorem 8.5.A. (Note that the additional automorphisms for the closed surfaces of genus 2 disappear when we pass to a subgroup of finite index not containing the hyperelliptic involution.) □

If S is a sphere with $\leqslant 4$ holes or a torus with $\leqslant 1$ holes, then Mod_S is indeed arithmetic: if S is a torus with $\leqslant 1$ holes, then Mod_S is isomorphic to $\mathrm{SL}_2(\mathbf{Z})$; if S is a sphere with 4 holes, then Mod_S is commensurable with $\mathrm{PSL}_2(\mathbf{Z}) = \mathrm{SL}_2(\mathbf{Z})/\{\pm I\}$ (cf., [101], Section 7 for a detailed discussion of this case); if S is a sphere with $\leqslant 3$ holes, then Mod_S is finite. If S is a torus with two holes, then Mod_S is not arithmetic. The above argument does not (currently) apply to this case, because Theorem 8.1.A does not include it. But older approaches, discussed in the next two sections (cf. 9.3 and 9.4) do apply to this case also.

It is worth stressing that this proof of the non-arithmeticity of the mapping class groups uses only very basic properties of arithmetic groups, going not much further than the definition, and a deep result (Theorem 8.1.A) from the topology of surfaces. This contrasts strongly with the original approach to the non-arithmeticity, which relied on the deepest properties of arithmetic groups; cf. 9.3.

9.3. *The original approach to the non-arithmeticity*

Arithmetic groups, like representations, can be divided into reducible and irreducible ones. An arithmetic group Γ is called *reducible* if it is "close" to a product of two infinite groups; in particular, if Γ is reducible, then it contains two subgroups Γ_1, Γ_2 of infinite index commuting with each other and generating together a subgroup of finite index in Γ. Using the last remark and Theorem 7.5.A, it is easy to see that no subgroup of finite index in $\mathrm{Mod}_S^\diamond$ can be isomorphic to a reducible arithmetic group. If Γ is not reducible, it is called *irreducible*.

By the definition, any arithmetic group Γ arises as a subgroup of a real semisimple Lie group H (without compact factors). The rank of the latter (i.e., the dimension of a Cartan subalgebra of its Lie algebra) is usually called the *rank* of Γ. In this section we will not address the question of whether this rank depends only on Γ or also on the realization of Γ as an arithmetic subgroup of some semisimple Lie group (cf. 9.4), and will consider our use of this notion as a slight abuse of language. The main point is that the properties of Γ depend strongly on whether the rank of Γ is 1 or $\geqslant 2$.

In particular, if Γ is an irreducible arithmetic group of rank $\geqslant 2$, then any normal subgroup of Γ is either a central finite subgroup or a subgroup of finite index. This is the famous Margulis finiteness theorem; see [246, Chapter 8] for the details. Now, for a closed surface S_g of genus g the kernel of the natural homomorphism $\mathrm{Mod}_{S_g} \to \mathrm{Sp}_{2g}(\mathbf{Z})$ from 9.1 is an infinite (and non-central) subgroup of infinite index. Indeed, this kernel, usually called the *Torelli subgroup* of Mod_{S_g}, contains, for example, all Dehn twists about separating circles and the above homomorphism is surjective, as we noted in 9.1. It follows that Mod_{S_g} cannot be an irreducible arithmetic group of rank $\geqslant 2$. Similar arguments apply to surfaces with non-empty boundary; for example, one can use the kernels of homomorphisms from 2.8 instead of the Torelli subgroup or a version of the latter.

It remains to show that Mod_S cannot be an irreducible arithmetic group of rank 1. One way to show this is to use the well known fact that any such arithmetic group contains a subgroup of finite index (actually, any torsion-free subgroup of finite index will work) which is isomorphic to the fundamental group of a complete Riemannian manifold of finite volume with curvature pinched between two negative constants, together with the following theorem.

THEOREM 9.3.A. *If S is not a sphere with $\leqslant 4$ holes and not a torus with $\leqslant 1$ holes, then no subgroup of finite index in Mod_S° is isomorphic to the fundamental group of a complete Riemannian manifold of finite volume with curvature pinched between two negative constants.*

PROOF. We only indicate the main idea. It is sufficient to prove that any subgroup of finite index in Mod_S contains a subgroup isomorphic to $(\mathbf{Z} * \mathbf{Z}) \times \mathbf{Z}$, and that none of the considered fundamental groups does. Subgroups of Mod_S isomorphic to $(\mathbf{Z} * \mathbf{Z}) \times \mathbf{Z}$ can be constructed in abundance with powers of Dehn twists as generators. The absence of such subgroups in the considered fundamental groups follows easily from the theory of Eberlein and O'Neill [48]; see [100], proof of Theorem 2 for the details. □

Irreducible arithmetic groups of rank 1 can also be excluded in a couple of other ways. See [79], Section 4.2 for a cohomological approach due to Harer, and still another approach due to Goldman. The latter is similar in spirit to the use of the Eberlein–O'Neill theory above.

The ideas outlined in this section, if combined with Mostow's rigidity theorem [184,185], lead also to some non-arithmeticity results for normal subgroups of Mod_S (even for so-called subnormal subgroups). For example, the Torelli subgroup, like the mapping class group itself, is not arithmetic. See [100].

Finishing the discussion of this original approach to the non-arithmeticity, we note that it uses almost no non-trivial information about mapping class groups, but relies heavily on the Margulis finiteness theorem, one of the deepest results about arithmetic groups at all. Cf. the remarks at the end of 9.2.

9.4. *Rank of the mapping class groups*

As we mentioned in 9.3, properties of an arithmetic group depend to a considerable extent on whether its rank is 1 or $\geqslant 2$. Given the analogy between mapping class groups and arithmetic groups, it is important to ascertain what arithmetic groups – of rank 1 or of rank $\geqslant 2$ – would be better analogues of mapping class groups. The best answer to this question would be a notion of rank applicable to both arithmetic groups and mapping class groups, together with a computation of this rank for the mapping class groups.

A suitable notion of rank was proposed by Ballmann and Eberlein [4], who developed the ideas of Prasad and Ragunathan [206]. Let Γ be an abstract group. Let $(\Gamma)_i$ denote the set of elements of Γ whose centralizer contains a free Abelian subgroup of rank $\leqslant i$ as a subgroup of finite index (so that $(\Gamma)_0 \subset (\Gamma)_1 \subset \cdots \subset (\Gamma)_i \subset \cdots$). Let $r(\Gamma)$ be the least natural number i such that Γ can be presented as a finite union

$$\Gamma = \gamma_1(\Gamma)_i \cup \cdots \cup \gamma_m(\Gamma)_i$$

of left translates of the set $(\Gamma)_i$ (the elements $\gamma_1, \ldots, \gamma_m \in \Gamma$ are arbitrary). Let us define the *rank* of Γ as the maximum of the numbers $r(\Gamma')$, where Γ' runs over all subgroups of finite index in Γ. We denote it by $\operatorname{rank}(\Gamma)$.

Prasad and Raghunathan [206] showed that if Γ is an arithmetic group, then $r(\Gamma)$ is equal to its rank in the sense of 9.3. In particular, this implies that the rank of an arithmetic group Γ in the sense of 9.3 depends only on the structure of Γ as an abstract group and not on its particular realization as an arithmetic subgroup of a semisimple Lie group. The replacement of $r(\cdot)$ by $\operatorname{rank}(\cdot)$ was suggested by Ballmann and Eberlein [4] in order to get a notion of rank invariant under passing to subgroups of finite index. They proved that if Γ is a discrete group of motions of a complete simply connected Riemannian manifold X whose curvature is nonpositive and bounded from below, such that X/Γ has finite volume, then $\operatorname{rank}(\Gamma)$ is equal to a geometrically defined rank of X. This result, in fact, includes the case of arithmetic groups, and means that this rank of an abstract group gives the desired result even for groups more general than arithmetic ones.

THEOREM 9.4.A. *If S is not a sphere with $\leqslant 3$ holes, then* $\operatorname{rank}(\operatorname{Mod}_S) = 1$.

PROOF. We will give only an outline of the proof. Let Γ be a torsion free subgroup of finite index in Mod_S (such subgroups exist in view of Corollary 7.1.B). Since $\operatorname{rank}(\cdot)$ is invariant under passing to subgroups of finite index, it is sufficient to show that $r(\Gamma) = 1$ for any such Γ. It is easy to see that $r(\Gamma) \neq 0$. Let Γ_{pA} denote the set of pseudo-Anosov elements of Γ. Theorem 7.5.A easily implies that the centralizers of pseudo-Anosov elements contain an infinite cyclic subgroup as a subgroup of finite index. Thus $\Gamma_{\text{pA}} \subset (\Gamma)_1$. Hence it suffices to prove that for some finite number of elements $g_1, \ldots, g_m \in \Gamma$ we have

$$\Gamma = g_1\Gamma_{\text{pA}} \cup \cdots \cup g_m\Gamma_{\text{pA}}.$$

In other words, it is sufficient to find a finite number of elements $g_1, \ldots, g_m$ such that for any $g \in \Gamma$ one of the elements $g_1^{-1}g, \ldots, g_m^{-1}g$ is pseudo-Anosov.

Using the compactness of the Thurston boundary of the Teichmüller space (cf. 7.2), it is possible to show that there exists a finite collection of isotopy classes $\alpha_1, \ldots, \alpha_n$ such that for any isotopy class β one of the pairs α_i, β fills S in the sense of 7.1. Fix such a collection $\alpha_1, \ldots, \alpha_n$. Let t_i, $1 \leqslant i \leqslant n$, be a non-trivial power of the Dehn twist t_{α_i} contained in Γ (remember that Γ is of finite index in Mod_S). Using Fathi's Theorem 7.1.D, one can show that for any $g \in \Gamma$ one of the elements $t_i^d g$, $1 \leqslant i \leqslant n$, $1 \leqslant d \leqslant 8$, is pseudo-Anosov. It follows that $r(\Gamma) = 1$. We refer to [102] for the details. □

COROLLARY 9.4.B. *For any compact orientable surface S, the group* Mod_S *is not isomorphic to any arithmetic group of rank* $\geqslant 2$.

One can replace all references to the Margulis finiteness theorem in 9.3 by this corollary. This leads to still another proof of the non-arithmeticity of the mapping class groups. On the arithmetic groups side, it uses the results of Prasad and Raghunathan [206], which are not as difficult as the Margulis finiteness theorem. It also uses some quite non-trivial results on the mapping class groups side (Fathi's Theorem 7.1.D); cf. the remarks at the end of 9.2, 9.3.

The next theorem, as we will explain in a moment, also supports the idea that the mapping class groups are similar to the arithmetic groups of rank 1, as opposed to rank $\geqslant 2$.

THEOREM 9.4.C. *Suppose that S is not a sphere with $\leqslant 3$ holes. Let us fix a finite set of generators of* Mod_S *and let us denote by $|\cdot|_W$ the minimal word length with respect to these generators. If f is an element of infinite order in* Mod_S, *then there is a constant $c > 0$ such that*

$$|f^n|_W \geqslant c|n|$$

for all n.

It is easy to see that one can always replace f by a nontrivial power of f in the proof. In particular, we may assume that f is a pure element of Mod_S (see 7.1). Then there are two different cases to consider. If the diffeomorphism induced on the result of cutting S along some realization of $\sigma(f)$ is isotopic to a pseudo-Anosov diffeomorphism on at least one component, then the theorem follows from the results of Mosher [182]. Otherwise, f is a composition of Dehn twists about several disjoint circles. This case was recently dealt with by Minsky [166], in response to a question posed by the author.

Recall that an arithmetic group Γ arises as a subgroup of a real semisimple Lie group H. The group Γ is called *cocompact* if H/Γ is compact. The similarity of Mod_S with cocompact arithmetic groups can be easily excluded by cohomological reasons (as, for example, in [79, Section 4.2]; cf. 9.3). But, for noncompact arithmetic groups Γ of rank $\geqslant 2$, there are always elements $f \in \Gamma$ of infinite order such that the word length $|f^n|_W$ grows slower than linearly (in fact, logarithmically), according to a recent result of Lubotzky, Mozes and Raghunathan [144]. Therefore, Theorem 9.4.C supports the analogy between Mod_S and rank 1 arithmetic groups. See [166] for other results of Farb, Lubotzky and Minsky supporting this analogy.

9.5. *The Mostow–Margulis superrigidity*

The Margulis superrigidity theorem [149], (cf. also [150] or [246, Chapter 5]) extends the Mostow rigidity theorem [184,185]. It asserts, in particular, that any homomorphism $\Gamma_1 \to \Gamma_2$ between arithmetic subgroups Γ_1, Γ_2 of semisimple Lie groups H_1, H_2 respectively (without compact factors) extends to a homomorphism $H_1 \to H_2$ after, perhaps, passing to a subgroup of finite index in Γ_1, if $\operatorname{rank} \Gamma_1 \geqslant 2$. Note that the restriction $\operatorname{rank} \Gamma_1 \geqslant 2$ in this theorem is necessary. Here we meet one of the situations where arithmetic groups of rank 1 and rank $\geqslant 2$ behave differently.

The analogy between mapping class groups and arithmetic groups suggests that one may extend this theorem to include mapping class groups on the equal footing with arithmetic groups. At first sight, this is impossible, because there are no Lie groups naturally (or, to the best of our knowledge, otherwise) containing mapping class groups. But this simply means that there should be no homomorphisms, or better, taking into account the possible passage to a finite index subgroup in the Margulis theorem, that any homomorphism should have finite image. In addition, since mapping class groups have rank 1 and the source group in the Margulis theorem should have rank $\geqslant 2$, we are forced to consider only homomorphisms from arithmetic groups of rank $\geqslant 2$ to mapping class groups. Then it is natural to expect that any such homomorphism has finite image.

The first step in the direction of this conjecture was made in [95], and the conjecture itself was stated explicitly in [97]. Further steps were made in [100]. Recently, Kaimanovich and Masur [118] almost completely proved it. Following [118], let us call a subgroup of Mod_S *non-elementary* if it leaves no finite subset of the Thurston boundary of the corresponding Teichmüller space invariant.

THEOREM 9.5.A. *Any non-elementary subgroup of a mapping class group is not isomorphic to an arithmetic group of* rank $\geqslant 2$.

See [118], Theorem 2.4.1. The proof of Kaimanovich and Masur is based on a theory of random walks on Teichmüller spaces and mapping class groups developed by Masur [155] and Kaimanovich and Masur [118] and the results of Furstenberg [56,57] about random walks and harmonic functions on discrete subgroups of Lie groups ([57] provides an excellent introduction to this circle of ideas). More precisely, Furstenberg introduced a property of harmonic functions (which are closely related to the random walks) on discrete groups which holds for all arithmetic groups of rank $\geqslant 2$ and does not hold for (at least some) arithmetic groups of rank 1. Kaimanovich and Masur show that this property does not hold also for all non-elementary subgroups of mapping class groups; thus they cannot be arithmetic groups of rank $\geqslant 2$. Note that these results of Furstenberg (as also the results of Mostow) preceded and influenced Margulis' approach to his superrigidity theorem.

If combined with the Margulis finiteness theorem and the results of [108] (cf. 7.4) about subgroups of mapping class groups, Theorem 9.5.A leads to the following corollary, which proves the above conjecture.

COROLLARY 9.5.B. *Let $\varphi : \Gamma \to \mathrm{Mod}_S$ be a homomorphism from an arithmetic group Γ to a mapping class group Mod_S. If $\operatorname{rank} \Gamma \geqslant 2$, then the image of φ is finite.*

A proof of this corollary along similar lines was also found by Farb and Masur [51]. It seems that it is possible to prove this corollary more directly, avoiding the use of the Margulis finiteness theorem, of random walks (and, hence, of Furstenberg's results), and even of Teichmüller spaces (but one still needs their Thurston's boundaries). Such an approach would follow the Margulis proof of his superrigidity theorem. I hope to return to these ideas in a future paper.

Acknowledgement

I am very grateful to R. Sher for the careful reading of this paper.

References

[1] W. Abikoff, *The Real Analytic Theory of Teichmüller Space*, Lecture Notes in Math., Vol. 820, Springer (1980).
[2] D.V. Anosov, *Geodesic flows on compact Riemannian manifolds of negative curvature*, Proc. Steklov Math. Inst. **90** (1967); English transl. Amer. Math. Soc. (1969).
[3] R. Baer, *Isotopien auf Kurven von orientierbaren, geschlossenen Flächen*, J. Reine Angew. Math. **159** (1928), 101–116.
[4] W. Ballmann and P. Eberlein, *Fundamental groups of manifolds of nonpositive curvature*, J. Differential Geom. **25** (1) (1987), 1–22.
[5] H. Bass and R. Kulkarni, *Uniform tree lattices*, J. Amer. Math. Soc. **3** (4) (1990), 843–902.
[6] H. Bass and A. Lubotzky, *Automorphisms of groups and of schemes of finite type*, Israel J. Math. **44** (1) (1983), 1–22.
[7] M. Bauer, *Examples of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. **330** (1) (1992), 333–370.
[8] H. Behr, *Über die endliche Definierbarkeit verallgemeinerter Einheitgruppen*, J. Reine Angew. Math. **211** (1962), 123–135.
[9] D.J. Benson, *Specht modules and the cohomology of mapping class groups*, Mem. Amer. Math. Soc. **90** (443) (1991), 29–92.
[10] D.J. Benson and F.R. Cohen, *The* mod 2 *cohomology of the mapping class group for a surface of genus two*, Mem. Amer. Math. Soc. **90** (443) (1991), 93–104.
[11] D.J. Benson and P.H. Kropholler, *Cohomology of groups*, Handbook of Algebraic Topology, I.M. James, ed., Elsevier (1995), 917–950.
[12] P. Bergau and J. Mennicke, *Über topologishe Abbildungen der Bretzekflache vom Geschlecht* 2, Math. Z. **74** (1960), 414–435.
[13] R. Bieri and B. Eckmann, *Groups with homological duality generalizing the Poincaré duality*, Invent. Math. **20** (2) (1973), 103–124.

[14] J.S. Birman, *Mapping class groups and their relationship to braid groups*, Comm. Pure App. Math. **22** (1969), 213–238.

[15] J.S. Birman, *Abelian quotients of the mapping class group of a 2-manifold*, Bull. Amer. Math. Soc. **76** (1) (1970), 147–150. Errata: Bull. Amer. Math. Soc. **77** (3) (1971), 479.

[16] J.S. Birman, *Braids, Links and Mapping Class Groups (based on lecture notes by James Cannon)*, Ann. of Math. Stud., Vol. 82, Princeton University Press (1974).

[17] J.S. Birman, *The algebraic structure of surface mapping class groups*, Discrete Groups and Automorphic Functions, W.J. Harvey, ed., Academic Press (1977), 163–198.

[18] J.S. Birman, *Mapping class groups of surfaces*, Braids (Santa Cruz, 1986), J.S. Birman and A. Libgober, eds, Contemp. Math. **78** (1988), 13–43.

[19] J.S. Birman and H.M. Hilden, *On mapping class groups of closed surfaces as covering spaces*, Advances in the Theory of Riemann Surfaces, L.V. Ahlfors, L. Bers, H.M. Farkas, R.C. Gunning, I. Kra and H.E. Rauch, eds, Ann. Math. Stud., Vol. 66, Princeton University Press (1971), 81–115.

[20] J.S. Birman, A. Lubotzky and J. McCarthy, *Abelian and solvable subgroups of the mapping class group*, Duke Math. J. **50** (4) (1983), 1107–1120.

[21] J.S. Birman and B. Wajnryb, *Errata: presentations of the mapping class group*, Israel J. Math. **88** (1–3) (1994), 425–427.

[22] C.-F. Bödigheimer, *On the topology of moduli spaces of Riemann surfaces. Part II: Homology operations*, Preprint, Mathematica Gottingensis, Heft 23 (1990).

[23] A. Borel and J.-P. Serre, *Corners and arithmetic groups*, Comment. Math. Helv. **48** (4) (1973), 436–491.

[24] B.H. Bowditch and D.B.A. Epstein, *Natural triangulations associated to a surface*, Topology **27** (1) (1988), 91–117.

[25] G.E. Bredon, *Geometry and Topology*, Graduate Texts in Math., Vol. 139, Springer (1993).

[26] K.S. Brown, *Cohomology of Groups*, Graduate Texts in Math., Vol. 87, Springer (1982).

[27] K.S. Brown, *Presentations for groups acting on simply-connected complexes*, J. Pure Appl. Algebra **32** (1) (1984), 1–10.

[28] K.S. Brown, *Buildings*, Springer (1989).

[29] P. Buser, *Geometry and Spectra of Compact Riemann Surfaces*, Progress in Math., Vol. 106, Birkhäuser (1992).

[30] A.J. Casson and S.A. Bleiler, *Automorphisms of Surfaces after Nielsen and Thurston*, London Math. Soc. Student Texts, Vol. 9, Cambridge Univ. Press (1989).

[31] J. Cerf, *Topologie de certaines espaces de plongements*, Bull. Soc. Math. France **89** (1961), 227–380.

[32] J. Cerf, *Sur les diffeomorphisms de la sphère de dimension trois* ($\Gamma_4 = 0$), Lecture Notes in Math., Vol. 53, Springer (1968).

[33] J. Cerf, *La stratification naturelle des espaces de functions différentiables rélles et le théorème de la pseudo-isotopie*, Publ. Math. IHES **39** (1970), 5–173.

[34] R. Charney and F. Cohen, *A stable splitting for the mapping class group*, Michigan Math. J. **35** (2) (1988), 269–284.

[35] R. Charney and R. Lee, *Characteristic classes for the classifying spaces of Hodge structures*, K-Theory **1** (3) (1987), 237–270.

[36] R. Charney and R. Lee, *An application of homotopy theory to mapping class groups*, J. Pure Appl. Algebra **44** (1–3) (1987), 127–135.

[37] F.R. Cohen, *Homology of the mapping class groups for surfaces of low genus*, The Lefschetz Centennial Conference, Part II. Proceedings on Algebraic Topology, S. Gitler, ed., Contemp. Math., Vol. 58 (1987), 21–30.

[38] F.R. Cohen, *Artin's braid group and the homology of certain subgroups of mapping class groups*, Mem. Amer. Math. Soc. **90** (443) (1991), 6–28.

[39] F.R. Cohen, *On the mapping class groups for punctured spheres, the hyperelliptic mapping class groups,* SO(3), *and* Spin^c(3), Amer. J. Math. **115** (2) (1993), 389–434.

[40] F.R. Cohen, *Mapping class groups and classical homotopy theory*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, C.-F. Bödigheimer and R. Hain, eds, Contemp. Math., Vol. 150 (1993), 63–74.

[41] F.R. Cohen, *On mapping class groups from a homotopy theoretic point of view*, Algebraic Topology: New Trends in Localization and Periodicity, C. Broto, C. Casacuberta and G. Mislin, eds, Progress in Math., Vol. 136, Birkhäuser (1996), 119–142.
[42] F.R. Cohen and U. Tillmann, *Toward homology operations for mapping class groups*, Homotopy Theory via Algebraic Geometry and Group Representations (Evanston, IL, 1997), M. Mahowald and S. Priddy, eds, Contemp. Math., Vol. 220 (1998), 35–46.
[43] M. Dehn, *Papers in Group Theory and Topology*, Translated and Introduced by John Stillwell, Springer (1987).
[44] M. Dehn, *On curve systems on two-sided surfaces, with application to the mapping problem*, Lecture (supplemented) to the math. colloquium, Breslau 11-2-1922, In: [43], 234–252.
[45] M. Dehn, *Die Gruppe der Abbildungsklassen*, Acta Math. **69** (1938), 135–206; English transl.: [43], 256–362.
[46] C.J. Earle and J. Eells, *A fibre bundle description of Teichmüller theory*, J. Differential Geom. **3** (1) (1969), 19–43.
[47] C.J. Earle and I. Kra, *On isometries between Teichmüller spaces*, Duke Math. J. **41** (3) (1974), 583–591.
[48] P. Eberlein and B. O'Neill, *Visibility manifolds*, Pacific J. Math. **46** (1) (1973), 45–109.
[49] D.B.A. Epstein, J.W. Cannon, D.F. Holt, S.V.F. Levy, M.S. Paterson and W.P. Thurston, *Word Processing in Groups*, Jones and Barlett Publishers (1992).
[50] D.B.A. Epstein and R.C. Penner, *Euclidean decompositions of noncompact hyperbolic manifolds*, J. Differential Geom. **27** (1) (1988), 67–80.
[51] B. Farb and H. Masur, *Superrigidity and mapping class groups*, Topology **37** (6) (1998), 1169–1176.
[52] A. Fathi, *Dehn twists and pseudo-Anosov diffeomorphisms*, Invent. Math. **87** (1) (1987), 129–151.
[53] A. Fathi, F. Laudenbach and V. Poénaru, *Travaux de Thurston sur les surfaces, Séminaire Orsay*, Astérisque, Vol. 66–67, Soc. Math. de France (1979).
[54] W. Fenchel, *Bemaerkinger om endlige grupper af afbildungklasser*, Mat. Tidsskr. B (1950), 90–95.
[55] R. Fricke and F. Klein, *Vorlesungen über die Theorie der automorphen Funktionen. I, Die gruppentheoretischen Grundlagen*, Teubner, Leipzig (1897).
[56] H. Furstenberg, *Poisson boundaries and envelopes of discrete groups*, Bull. Amer. Math. Soc. **73** (3) (1967), 350–356.
[57] H. Furstenberg, *Random walks and discrete subgroups of Lie groups*, Advances in Probability and Related Topics, Vol. 1, Marcel Dekker (1971), 3–63.
[58] F.P. Gardiner, *Teichmüller Theory and Quadratic Differentials*, Wiley (1987).
[59] S. Gervais, *Presentation and central extension of mapping class groups*, Trans. Amer. Math. Soc. **348** (8) (1996), 3097–3132.
[60] S. Gervais, *A finite presentation of the mapping class group of an oriented surface*, Preprint (1998), 22 pp.
[61] J. Gilman, *Structures of elliptic irreducible subgroups of the modular group*, Proc. London. Math. Soc. **47** (1) (1983), 27–42.
[62] H. Glover and G. Mislin, *On the stable cohomology of the mapping class groups*, Algebraic Topology, Göttingen 1984, L. Smith, ed., Lecture Notes in Math., Vol. 1172 (1985), 80–84.
[63] H. Glover and G. Mislin, *Torsion in the mapping class group and its cohomology*, J. Pure Appl. Algebra **44** (1–3) (1987), 177–189.
[64] H.H. Glover, G. Mislin and Y. Xia, *On the Farrel cohomology of mapping class groups*, Invent. Math. **105** (3) (1992), 535–545.
[65] H.H. Glover, G. Mislin and Y. Xia, *On the Yagita invariant of mapping class groups*, Topology **33** (3) (1994), 557–574.
[66] L. Greenberg, *Maximal groups and Riemann surfaces*, Discontinuous Groups and Riemann Surfaces, L. Greenberg, ed., Ann. of Math. Stud., Vol. 79, Princeton University Press (1974), 207–226.
[67] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, S.M. Gersten, ed., MSRI Publications, Vol. 8, Springer (1987), 75–263.
[68] E. Grossman, *On the residually finiteness of certain mapping class groups*, J. London Math. Soc. **9** (1) (1974), 160–164.
[69] R.M. Hain, *Completions of mapping class groups and the cycle $C - C^{-}$*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, C.-F. Bödigheimer and R. Hain, eds, Contemp. Math., Vol. 150 (1993), 75–105.

[70] R.M. Hain, *Torelli groups and geometry of moduli spaces of curves*, Current Topics in Algebraic Geometry (Berkeley, CA, 1992/93), MSRI Publications, Vol. 28, Cambridge University Press (1995), 97–143.
[71] R.M. Hain, *Infinitesimal presentations of the Torelli groups*, J. Amer. Math. Soc. **10** (3) (1997), 597–651.
[72] R.M. Hain and E. Looijenga, *Mapping class groups and moduli spaces of curves*, Algebraic Geometry – Santa Cruz 1995, Proc. Sympos. Pure Math., Vol. 62, Part 2, Amer. Math. Soc. (1997), 97–142.
[73] M.-E. Hamstrom, *The space of homeomorphisms of a torus*, Illinois J. Math. **9** (1965), 59–65.
[74] M.-E. Hamstrom, *Homotopy groups of the space of homeomorphisms of a 2-manifold*, Illinois J. Math. **10** (1966), 563–573.
[75] J.L. Harer, *The second homology group of the mapping class group of an orientable surface*, Invent. Math. **72** (2) (1983), 221–239.
[76] J.L. Harer, *The homology of the mapping class group and its connection to surface bundles over surfaces*, Four-Manifold Theory, C. Gordon and R. Kirby, eds, Contemp. Math., Vol. 35 (1984), 311–314.
[77] J.L. Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. of Math. **121** (2) (1985), 215–249.
[78] J.L. Harer, *The virtual cohomology dimension of the mapping class group of an orientable surface*, Invent. Math. **84** (1) (1986), 157–176.
[79] J.L. Harer, *The cohomology of the moduli space of curves*, Lecture Notes in Math., Vol. 1337, Springer (1988), 138–221.
[80] J.L. Harer, *Stability of the homology of the moduli spaces of Riemann surfaces with spin structure*, Math. Ann. **287** (2) (1990), 323–334.
[81] J.L. Harer, *The third homology group of the moduli space of curves*, Duke Math. J. **63** (1) (1991), 25–55.
[82] J.L. Harer, *The rational Picard group of the moduli space of Riemann surfaces with spin structure*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, C.-F. Bödigheimer and R. Hain, eds, Contemp. Math., Vol. 150 (1993), 107–136.
[83] J.L. Harer, *The fourth homology group of the moduli space of curves*, Preprint (1993), 36 pp.
[84] J.L. Harer, *Improved homology stability for the homology of the mapping class groups of surfaces*, Preprint (1993), 18 pp.
[85] J.L. Harer and D. Zagier, *The Euler characteristic of the moduli space of curves*, Invent. Math. **85** (3) (1986), 457–485.
[86] W.J. Harvey, *Geometric structure of surface mapping-class groups*, Homological Group Theory, C.T.C. Wall, ed., London Math. Soc. Lecture Notes Series, Vol. 36, Cambridge University Press (1979), 255–269.
[87] W.J. Harvey, *Boundary structure of the modular group*, Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference, I. Kra and B. Maskit, eds, Ann. of Math. Stud., Vol. 97, Princeton University Press (1981), 245–251.
[88] A. Hatcher, *Homeomorphisms of sufficiently large P^2-irreducible 3-manifolds*, Topology **15** (4) (1976), 343–347.
[89] A. Hatcher, *On triangulations of surfaces*, Topology Appl. **40** (2) (1991), 189–194.
[90] A. Hatcher, *Measured lamination spaces for surfaces, from the topological viewpoint*, Topology Appl. **30** (1) (1998), 63–88.
[91] A. Hatcher and W. Thurston, *A presentation for the mapping class group of a closed orientable surface*, Topology **19** (3) (1980), 221–237.
[92] J. Hemion, *On the classification of homeomorphisms of 2-manifolds and the classification of 3-manifolds*, Acta Math. **142** (1–2) (1979), 123–155.
[93] S.P. Humphries, *Generators for the mapping class group*, Topology of Low-Dimensional Manifolds (Proc. Second Sussex Conf., Chelwood Gate, 1977), Lecture Notes in Math., Vol. 722 (1979), 44–47.
[94] N.V. Ivanov, *Groups of diffeomorphisms of Waldhausen manifolds*, Studies in Topology II, Notes of LOMI Scientific Seminars, Vol. 66 (1976), 172–176; English transl.: J. Soviet Math. **12** (1) (1979), 115–118.
[95] N.V. Ivanov, *Algebraic properties of the Teichmüller modular group*, Dokl. Akad. Nauk SSSR **275** (4) (1984), 786–789; English transl.: Soviet Math. Dokl. **29** (2) (1984), 288–291.
[96] N.V. Ivanov, *On the virtual cohomology dimension of the Teichmüller modular group*, Lecture Notes in Math., Vol. 1060, Springer (1984), 306–318.
[97] N.V. Ivanov, *Algebraic properties of mapping class groups of surfaces*, Banach Center Publications, Vol. 18, PWN – Polish Scientific Publishers (1986), 15–35.

[98] N.V. Ivanov, *Complexes of curves and the Teichmüller modular group*, Uspekhi Mat. Nauk **42** (3) (1987), 49–91; English transl.: Russian Math. Surveys **42** (3) (1987), 55–107.
[99] N.V. Ivanov, *Subgroups of Teichmüller modular groups and their Frattini subgroups*, Funktsional. Anal. i Prilozhen. **21** (2) (1987), 76–77; English transl.: Functional Anal. Appl. **21** (2) (1987), 154–155.
[100] N.V. Ivanov, *Teichmüller modular groups and arithmetic groups*, Research in Topology 6, Notes of LOMI Scientific Seminars, Vol. 167 (1988), 95–110; English transl.: J. Soviet Math. **52** (1) (1990), 2809–2818.
[101] N.V. Ivanov, *Automorphisms of Teichmüller modular groups*, Lecture Notes in Math., Vol. 1346, Springer (1988), 199–270.
[102] N.V. Ivanov, *Rank of Teichmüller modular groups*, Mat. Zametki **44** (5) (1988), 636–644; English transl.: Math. Notes **44** (5–6) (1988), 829–832.
[103] N.V. Ivanov, *Automorphisms of complexes of curves and of Teichmüller spaces*, Preprint IHES/M/89/60, 1989, 13 pp.; Also in: Progress in Knot Theory and Related Topics, Travaux en Cours, Vol. 56, Hermann, Paris (1997), 113–120.
[104] N.V. Ivanov, *Stabilization of the homology of Teichmüller modular groups*, Algebra i Analiz **1** (3) (1989), 110–126; English transl.: Leningrad J. of Math. **1** (3) (1990), 675–691.
[105] N.V. Ivanov, *Attaching corners to Teichmüller space*, Algebra i Analiz **1** (5) (1989), 115–143. English transl.: Leningrad J. of Math. **1** (5) (1990), 1177–1205.
[106] N.V. Ivanov, *Complexes of curves and Teichmüller spaces*, Mat. Zametki **49** (5) (1991), 54–61; English transl.: Math. Notes **49** (5–6) (1991), 479–484.
[107] N.V. Ivanov, *On the finite approximability of modular Teichmüller groups*, Sibirsk. Mat. Zh. **32** (1) (1991), 182–185; English transl.: Siberian Math. J. **32** (1) (1991), 148–150.
[108] N.V. Ivanov, *Subgroups of Teichmüller Modular Groups*, Transl. Math. Monographs, Vol. 115, Amer. Math. Soc. (1992).
[109] N.V. Ivanov, *On the homology stability for Teichmüller modular groups: Closed surfaces and twisted coefficients*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, C.-F. Bödigheimer and R. Hain, eds, Contemp. Math., Vol. 150 (1993), 149–194.
[110] N.V. Ivanov, *Automorphisms of complexes of curves and of Teichmüller spaces*, Internat. Math. Res. Notices **14** (1997), 651–666.
[111] N.V. Ivanov and J.D. McCarthy, *On injective homomorphisms between Teichmüller modular groups*, Preprint (1995), 85 pp.
[112] N.V. Ivanov and J.D. McCarthy, *On injective homomorphisms between Teichmüller modular groups I*, Invent. Math., **135** (2) (1999), 425–486.
[113] D.L. Johnson, *Homeomorphisms of a surface which act trivially on homology*, Proc. Amer. Math. Soc. **75** (1) (1979), 119–125.
[114] D.L. Johnson, *A survey of the Torelli group*, Low-Dimensional Topology (San Francisco, Calif., 1981), Contemp. Math., Vol. 20 (1983), 165–179.
[115] D.L. Johnson, *The structure of the Torelli group. I. A finite set of generators for $\mathcal{I}$*, Ann. of Math. **118** (3) (1983), 423–442.
[116] D.L. Johnson, *The structure of the Torelli group. II. A characterization of the group generated by twists on bounding curves*, Topology **24** (2) (1985), 113–126.
[117] D.L. Johnson, *The structure of the Torelli group. III. The abelianizaton of $\mathcal{T}$*, Topology **24** (2) (1985), 127–144.
[118] V.A. Kaimanovich and H. Masur, *The Poisson boundary of the mapping class group*, Invent. Math. **125** (2) (1996), 221–264.
[119] W. van der Kallen, *Homology stability for linear groups*, Invent. Math. **60** (3) (1980), 269–295.
[120] M. Kapovich, *Hyperbolic Manifolds and Discrete Groups*, Progress in Math, Vol. 183, Birkhäuser, 2001.
[121] S.P. Kerckhoff, *The asymptotic geometry of Teichmüller spaces*, Topology **19** (1) (1980), 23–41.
[122] S.P. Kerckhoff, *The Nielsen realization problem*, Bull. Amer. Math. Soc. **2** (3) (1980), 452–454.
[123] S.P. Kerckhoff, *The Nielsen realization problem*, Ann. of Math. **117** (2) (1983), 235–265.
[124] S.P. Kerckhoff, *The geometry of Teichmüller space*, Proc. of the International Congress of Mathematicians, Vols. 1, 2 (Warsaw, 1983), PWN – Polish Scientific Publishers, Warsaw (1984), 665–678.
[125] S.P. Kerckhoff, *Lines of minima in Teichmüller space*, Duke Math. J. **65** (2) (1992), 187–213.
[126] M. Kontsevich, *Intersection theory on the moduli space of curves and the matrix Airy function*, Comm. Math. Phys. **147** (1) (1992), 1–23.

[127] M. Korkmaz, *Complexes of curves and mapping class groups*, PhD Thesis, Michigan State University (1996); Also in: M. Korkmaz, *Automorphisms of complexes of curves on punctured spheres and on punctured tori*, Topology Appl. **95** (2) (1999), 85–111.

[128] A. Kosinsky, *Differential Manifolds*, Pure and Appl. Math., Vol. 138, Academic Press (1993).

[129] J.L. Koszul, *Lectures on groups of transformations*, Notes by R.R. Simha and R. Sridharan, Tata Institute of Fundamental Research, Bombay (1965).

[130] F. Laudenbach, *Presentation du groupe de diffeotopies d'une surface compacte orientable*, In: [53], Exp. 15, 267–282.

[131] R. Lee and S. Weintraub, *Cohomology of* $Sp(4, \mathbf{Z})$ *and the related groups and spaces*, Topology **24** (4) (1985), 391–410.

[132] W.B.R. Lickorish, *A representation of orientable combinatorial* 3*-manifolds*, Ann. of. Math. **76** (3) (1962), 531–540.

[133] W.B.R. Lickorish, *Homeomorphisms of non-orientable two-manifolds*, Proc. Cambridge Phil. Soc. **59** (2) (1963), 307–317.

[134] W.B.R. Lickorish, *A finite set of generators for the homeotopy group of a* 2*-manifold*, Proc. Cambridge Phil. Soc. **60** (4) (1964), 769–778.

[135] W.B.R. Lickorish, *Homeomorphisms of non-orientable two-manifolds*, Proc. Cambridge Phil. Soc. **61** (1965), 61–64.

[136] W.B.R. Lickorish, *A foliation of* 3*-manifolds*, Ann. of. Math. **82** (3) (1965), 414–420.

[137] W.B.R. Lickorish, *Corrigendum: On the homeotopy group of a* 2*-manifold*, Proc. Cambridge Phil. Soc. **62** (4) (1966), 679–681.

[138] D.D. Long, *Constructing pseudo-Anosov maps*, Knot Theory and Manifolds (Vancouver, BC, 1983), Lecture Notes in Math., Vol. 1144 (1985), 108–114.

[139] D.D. Long, *A note on the normal subgroups of mapping class groups*, Math. Proc. Cambridge Philos. Soc. **99** (1) (1986), 79–87.

[140] D.D. Long and H. Morton, *Hyperbolic* 3*-manifolds and surface homeomorphisms*, Topology **25** (4) (1986), 575–583.

[141] E. Looijenga, *Intersection theory on Deligne–Mumford compactifications [affter Witten and Kontsevich]*, Séminaire Bourbaki, Vol. 1992/1993, Exp. 768, Astérisque, Vol. 216 (1993), 187–212.

[142] E. Looijenga, *Cellular decompositions of compactified moduli spaces of pointed curves*, The Moduli Space of Curves, Progress in Math., Vol. 129, Birkhäuser (1995), 369–400.

[143] E. Looijenga, *Stable homology of the mapping class group with symplectic coefficients and the universal Abel–Jacobi map*, J. Algebraic Geom. **5** (1) (1996), 135–150.

[144] A. Lubotzky, S. Mozes and M.S. Raghunathan, *Cyclic subgroups of exponential growth and metric on discrete groups*, C. R. Acad. Sci. Paris **317** (8) (1993), 735–740.

[145] F. Luo, *On nonseparating simple closed curves in a compact surface*, Topology **36** (1997), 381–410.

[146] F. Luo, *A presentation of the mapping class group*, Math. Res. Lett. **4** (5) (1997), 735–739.

[147] F. Luo, *Automorphisms of complexes of curves*, Topology **39** (2) (2000), 283–298.

[148] W. Magnus, *Über Automorphismen von Fundamentalgruppen berandeter Flächen*, Math. Ann. **109** (1934), 617–648.

[149] G.A. Margulis, *Arithmeticity of the irreducible lattices in the semisimple groups of rank greater than* 1, Invent. Math. **76** (1) (1984), 93–120.

[150] G.A. Margulis, *Discrete Subgroups of Semisimple Lie Groups*, Ergebnisse der Math., 3 Folge, B. 17, Springer (1991).

[151] G.A. Margulis and G.A. Soifer, *Maximal subgroups of infinite index in finitely generated linear groups*, J. Algebra **69** (1) (1981), 1–23.

[152] H. Masur, *On a class of geodesics in Teichmüller space*, Ann of Math. **102** (2) (1975), 205–221.

[153] H. Masur, *Uniquely ergodic quadratic differentials*, Comment. Math. Helv. **55** (2) (1980), 255–266.

[154] H. Masur, *Interval exchange transformations and measured foliations*, Ann. of Math. **115** (1) (1982), 169–201.

[155] H. Masur, *Random walks on Teichmüller spaces and the mapping class group*, J. Anal. Math. **67** (1995), 117–164.

[156] H. Masur and Y. Minsky, *Geometry of the complexes of curves I: Hyperbolicity*, Invent. Math. **138** (1) (1999), 103–149.

[157] H. Masur and Y. Minsky, *Geometry of the complexes of curves I: Hierarchical structure*, Geom. Funct. Anal. **10** (4) (2000), 902–974.

[158] J.D. McCarthy, *Normalizers and centralizers of pseudo-Anosov mapping classes*, Preprint (1982).

[159] J.D. McCarthy, *A "Tits-alternative" for subgroups of surface mapping class groups*, Trans. Amer. Math. Soc. **291** (2) (1985), 583–612.

[160] J.D. McCarthy, *Automorphisms of surface mapping class groups. A recent theorem of N. Ivanov*, Invent. Math. **84** (1) (1986), 49–71.

[161] J. McCool, *Some finitely presented subgroups of the automorphism group of a free group*, J. Algebra **35** (1975), 205–213.

[162] C.T. McMullen, *Renormalization and 3-Manifolds which Fiber over the Circle*, Ann. Math. Stud., Vol. 142, Princeton University Press (1996).

[163] G. Mess, *Unit tangent bundle subgroups of the mapping class groups*, Preprint IHES/M/90/30 (1990), 15.

[164] E.Y. Miller, *The homology of the mapping class group*, J. Differential Geom. **24** (1) (1986), 1–14.

[165] Y.N. Minsky, *A geometric approach to the complex of curves on a surface*, Topology and Teichmüller Spaces (Katinkulta, 1995), World Sci. (1996), 149–159.

[166] Y.N. Minsky, *Dehn twists have linear growth*, Preprint (1997), 5 pp.; Also in: B. Farb, A. Lubotzky and Y. Minsky, *Rank-1 phenomena for mapping class groups*, Duke Math. J. **106** (3) (2001), 581–597.

[167] G. Mislin, *Mapping class groups, characteristic classes, and Bernoulli numbers*, The Hilton Symposium 1993. Topics in Topology and Group Theory, G. Mislin, ed., CRM Proceedings & Lecture Notes, Vol. 6, Amer. Math. Soc. (1994), 103–131.

[168] Sh. Morita, *Characteristic classes of surface bundles*, Bull. Amer. Math. Soc. **11** (2) (1984), 386–388.

[169] Sh. Morita, *Characteristic classes of surface bundles*, Invent. Math. **90** (3) (1987), 552–577.

[170] Sh. Morita, *Characteristic classes of surface bundles and bounded cohomology*, A Fête of Topology, Academic Press (1988), 233–257.

[171] Sh. Morita, *Families of Jacobian manifolds and characteristic classes of surface bundles. I*, Ann. Inst. Fourier, Grenoble **39** (3) (1989), 777–810.

[172] Sh. Morita, *Families of Jacobian manifolds and characteristic classes of surface bundles. II*, Math. Proc. Cambridge Phil. Soc. **105** (1) (1989), 79–101.

[173] Sh. Morita, *Casson's invariant for homology 3-spheres and characteristic classes of surface bundles. I*, Topology **28** (3) (1989), 305–323.

[174] Sh. Morita, *Mapping class groups of surfaces and three-dimensional manifolds*, Proceedings of the International Congress of Mathematicians (Kyoto, 1990), Vols. I, II, Math. Soc. Japan (1991), 665–674.

[175] Sh. Morita, *Characteristic classes of surface bundles and Casson invariant*, Sugaku **43** (3) (1991), 232–247; English transl.: Sugaku Expositions **7** (1) (1994), 59–79.

[176] Sh. Morita, *The structure of the mapping class group and characteristic classes of surface bundles*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, C.-F. Bödigheimer and R. Hain, eds, Contemp. Math., Vol. 150 (1993), 303–315.

[177] Sh. Morita, *The extension of Johnson's homomorphism from the Torelli group to the mapping class group*, Invent. Math. **111** (1) (1993), 197–224.

[178] Sh. Morita, *Abelian quotients of subgroups of the mapping class group of surfaces*, Duke Math. J. **70** (3) (1993), 699–726.

[179] Sh. Morita, *Casson invariant, signature defect of framed manifolds and the secondary characteristic classes of surface bundles*, J. Differential Geom. **47** (3) (1997), 560–599.

[180] L. Mosher, *The classification of pseudo-Anosovs*, Low Dimensional Topology and Kleinian Groups (Coventry/Durham, 1984), London Math. Soc. Lecture Notes Ser., Vol. 112, Cambridge Univ. Press (1986), 13–75.

[181] L. Mosher, *Mapping class groups are automatic*, Math. Res. Lett. **1** (2) (1994), 249–255.

[182] L. Mosher, *Mapping class groups are automatic*, Ann. of Math. **142** (2) (1995), 303–384.

[183] L. Mosher, *A user's guide to the mapping class group: once punctured surfaces*, Geometric and Computational Perspectives on Infinite Groups (Minneapolis, MN and New Brunswick, NJ, 1994), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., Vol. 25 (1996), 101–174.

[184] G.D. Mostow, *Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms*, Publ. Math. IHES **34** (1968), 53–104.

[185] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Ann. of Math. Stud., Vol. 78, Princeton University Press (1973).
[186] D. Mumford, *Abelian quotients of the Teichmüller modular group*, J. Anal. Math. **18** (1967), 227–244.
[187] D. Mumford, *A remark on Mahler's compactness theorem*, Proc. Amer. Math. Soc. **28** (1) (1971), 289–294.
[188] D. Mumford, *Towards an enumerative geometry of the moduli space of curves*, Arithmetic and Geometry, Vol. II, Progress in Math., Vol. 36, Birkhäuser (1983), 271–328.
[189] M. Näätänen and R.C. Penner, *The convex hull construction for compact surfaces and the Dirichlet polygon*, Bull. London Math. Soc. **23** (6) (1991), 568–574.
[190] W.D. Neumann and A.W. Reid, *Arithmetic of hyperbolic manifolds*, Topology '90, B. Apanasov, W.D. Neumann, A.W. Reid and L. Siebenmann, eds, Walter de Gruyter (1992), 273–310.
[191] J. Nielsen, *Untersuchungen zur Theorie der geschlossenen zweiseitigen Flächen I*, Acta Math. **50** (1927), 189–358.
[192] J. Nielsen, *Untersuchungen zur Theorie der geschlossenen zweiseitigen Flächen II*, Acta Math. **53** (1929), 1–76.
[193] J. Nielsen, *Untersuchungen zur Theorie der geschlossenen zweiseitigen Flächen III*, Acta Math. **58** (1932), 87–167.
[194] J. Nielsen, *Abbildungsklassen der endlicher Ordnung*, Acta Math. **75** (1942), 23–115.
[195] J.-P. Otal, *Le théorème d'hyperbolisation pour les variétés fibréss de dimension* 3, Astérisque, Vol. 235, Soc. Math. de France (1996).
[196] R. Palais, *Local triviality of the restriction map for embeddings*, Comment. Math. Helv. **34** (1960), 305–312.
[197] R.C. Penner, *The decorated Teichmüller space of punctured surfaces*, Comm. Math. Phys. **113** (2) (1987), 299–339.
[198] R.C. Penner, *The moduli space of a punctured surface and perturbative series*, Bull. Amer. Math. Soc. **15** (1) (1986), 73–77.
[199] R.C. Penner, *Perturbative series and the moduli space of Riemann surfaces*, J. Differential Geom. **27** (1) (1988), 35–53.
[200] R.C. Penner, *A construction of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. **310** (1) (1988), 179–197.
[201] R.C. Penner, *Weil–Petersson volumes*, J. Differential Geom. **35** (3) (1991), 559–608.
[202] R.C. Penner, *Universal constructions in Teichmüller theory*, Adv. in Math. **98** (2) (1993), 143–215.
[203] V.P. Platonov, *Frattini subgroups of linear groups and finitary approximability*, Dokl. Akad. Nauk SSSR **171** (1966), 798–801; English transl.: Soviet Math. Dokl. **7** (6) (1966), 1557–1560.
[204] T. Poston and I.N. Stewart, *Taylor expansions and catastrophes*, Pitman Res. Notes Math., Vol. 7 (1976).
[205] J. Powell, *Two theorems on the mapping class group of a surface*, Proc. Amer. Math. Soc. **68** (3) (1978), 347–350.
[206] G. Prasad and M.S. Raghunathan, *Cartan subgroups and lattices in semisimple groups*, Ann. of Math. **96** (2) (1972), 296–317.
[207] S. Priddy and Y. Xia, *Cohomological invariants of mapping class groups*, Quart. J. Math., Oxford **47** (187) (1996), 361–370.
[208] H.L. Royden, *Automorphisms and isometries of Teichmüller spaces*, Advances in the Theory of Riemann Surfaces, L.V. Ahlfors, L. Bers, H.M. Farkas, R.C. Gunning, I. Kra, H.E. Rauch, eds, Ann. of Math. Stud., Vol. 66, Princeton University Press (1971), 369–383.
[209] P. Schmutz, *Die Parametrisierung des Teichmüllerraumes durch geodätische Längenfunktionen*, Comment. Math. Helv. **68** (2) (1993), 278–288.
[210] G.P. Scott, *The space of homeomorphisms of a* 2*-manifold*, Topology **9** (1970), 97–109.
[211] H. Seifert, *Bemerkungen zur stetigen Abbildungen von Flächen*, Abh. Math. Sem. Univ. Hamburg **12** (1937), 23–37.
[212] J.-P. Serre, *Rigiditi de foncteur d'Jacobi d'échelon* $n \geqslant 3$, Sem. H. Cartan (1960/1961), Appendix to Exp. 17.
[213] J.-P. Serre, *Cohomologie des groupes discretes*, Prospects in Mathematics, Ann. of Math. Stud., Vol. 70, Princeton University Press (1971), 77–169.
[214] J.-P. Serre, *Arbres, amalgames*, SL_2, rédigé avec la collaboration de Hyman Bass, Astérisque, Vol. 46 (1977); English transl.: *Trees*, translated by J. Stillwell, Springer (1980).

[215] S. Smale, *Diffeomorphisms of the 2-sphere*, Proc. Amer. Math. Soc. **10** (1959), 621–626.
[216] E. Spanier, *Algebraic Topology*, 2nd ed., Springer (1990).
[217] J. Stillwell, *Geometry of Surfaces*, Universitexts, Springer (1992).
[218] K. Strebel, *On quadratic differentials with closed trajectories and second order poles*, J. Anal. Math. **19** (1967), 373–382.
[219] K. Strebel, *Quadratic Differentials*, Ergebnisse der Math., 3 Folge, Band 5, Springer (1984).
[220] D. Sullivan, *Travaux de Thurston sur les groupes quasi-fuchsiens et sur les variétés hyperboliques de dimension 3 fibrées sur S^1*, Séminaire Bourbaki, Vol. 1979/80, Exp. 554, Lecture Notes in Math., Vol. 842, Springer (1981), 196–214.
[221] R.T. Tchangang, *Le groupe d'automorphismes du groupe modulaire*, Ann. Inst. Fourier (Grenoble) **37** (2) (1987), 19–31.
[222] W.P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. **19** (2) (1988), 417–431.
[223] W.P. Thurston, *Hyperbolic structures on 3-manifolds II: Surface groups and 3-manifolds which fiber over the circles*, Preprint (1986).
[224] U. Tillmann, *Discrete models for the category of Riemann surfaces*, Math. Proc. Cambridge Phil. Soc. **121** (1) (1997), 39–49.
[225] U. Tillmann, *On the homotopy of the stable mapping class group*, Invent. Math. **130** (1997), 257–275.
[226] J. Tits, *Free subgroups in linear groups*, J. Algebra **20** (1972), 250–270.
[227] J. Tits, *Buildings of Spherical Type and Finite BN-pairs*, Lecture Notes in Math., Vol. 386, Springer (1974).
[228] A.J. Tromba, *Teichmüller Theory in Riemannian Geometry (based on notes by J. Dezler)*, Lectures in Math. ETH Zürich, Birkhäuser (1992).
[229] A.J. Tromba, *Dirichlet's energy on Teichmüller moduli space and the Nielsen realization problem*, Math. Z. **222** (3) (1996), 451–464.
[230] B. Wajnryb, *A simple presentation for the mapping class group of an orientable surface*, Israel J. Math. **45** (2–3) (1983), 157–174.
[231] B. Wajnryb, *Mapping class group of a surface is generated by two elements*, Topology **35** (2) (1996), 377–383.
[232] B. Wajnryb, *An elementary approach to the mapping class group of a surface*, Geom. Topology **3** (1999), 405–466 (electronic).
[233] E. Witten, *Two dimensional gravity and intersection theory on moduli spaces*, Surveys in Differential Geometry, Vol. 1 (1991), 243–310.
[234] S.A. Wolpert, *Geodesic length functionals and the Nielsen problem*, J. Differential Geom. **25** (1987), 275–295.
[235] Y. Xia, *The p-period of an infinite group*, Publ. Mat. **36** (1) (1992), 241–250.
[236] Y. Xia, *The p-periodicity of the mapping class group and an estimate of its p-period*, Proc. Amer. Math. Soc. **116** (4) (1992), 1161–1169.
[237] Y. Xia, *The p-torsion of the Farrel–Tate cohomology of the mapping class group $\Gamma_{(p-1)/2}$*, Topology '90, B. Apanasov, W.D. Neumann, A.W. Reid and L. Siebenmann, eds, Walter de Gruyter (1992), 391–398.
[238] Y. Xia, *The p-torsion of the Farrell–Tate cohomology of the mapping class group Γ_{p-1}*, J. Pure Appl. Algebra **78** (3) (1992), 319–334.
[239] Y. Xia, *An obstruction in the mapping class group*, Bull. London Math. Soc. **27** (1) (1995), 51–57.
[240] Y. Xia, *On the cohomology of Γ_p*, Trans. Amer. Math. Soc. **347** (9) (1995), 3659–3670.
[241] Y. Xia, *On the connectivity of posets in the mapping class groups*, Algebraic Topology: New Trends in Localization and Periodicity, C. Broto, C. Casacuberta and G. Mislin, eds, Progress in Math., Vol. 136, Birkhäuser (1996), 393–400.
[242] Y. Xia, *Geodesic length functions and cohomology of Γ_p*, Preprint (1995), 14 pp.
[243] D. Zagier, *On the distribution of the number of cycles of elements in symmetric groups*, Nieuw Archief voor Wiskunde **13** (3) (1995), 489–495.
[244] H. Zieschang, *Discrete groups of motions of the plane and plane group diagrams*, Uspekhi Mat. Nauk **29** (3) (1966), 195–212 (in Russian).
[245] H. Zieschang, E. Vogt and H.-D. Coldeway, *Surfaces and Planar Discontinuous Groups*, Lecture Notes in Math., Vol. 835, Springer (1980).
[246] R.J. Zimmer, *Ergodic Theory and Semisimple Groups*, Monographs Math., Vol. 81, Birkhäuser (1984).

CHAPTER 13

Seifert Manifolds

Kyung Bai Lee

University of Oklahoma, Norman, OK 73019, USA

Frank Raymond

University of Michigan, Ann Arbor, MI 48109, USA

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

Seifert fibered spaces have played an interesting and important role in topology and transformation groups. The classical definition of Seifert [60] in 1933, concerning circle bundles with specified singularities over 2-dimensional manifolds, preceded much of the notions and development of bundle theory. In fact, Seifert's splendid paper had a profound influence on both bundle theory and low-dimensional topology. Recognition of the importance of Seifert fiber spaces to transformation groups was communicated a long time ago to one of the authors by Deane Montgomery.

Many of the interesting spaces in topology and geometry admit the structure of a bundle with singularities. The geometric structure is usually displayed as part of this bundle structure. While Seifert fiberings, in their utmost generality, could be said to encompass all of these phenomena, we have concentrated here, in our exposition, on a particular type called *injective Seifert fiberings*. These seem to entail the simplest structure in that they admit a global uniformization, have close connections with transformation groups, and have many applications. Moreover, because topologists have become intrigued with the "geometrization" of manifolds, we hope that this paper will also inform the reader of the significant role that Seifert fiber spaces and Seifert constructions can play in these matters.

We have tried to make our exposition as concrete and accessible as possible. This is especially true in the beginning and in the last two sections where examples and illustrations of the theory are worked out in detail. Each section has its own introduction and it will be helpful to read Sections 2, 5 and 6 first. To digest Sections 3 and 4, it is advisable to read these through first for an overview before checking the details.

Notation. We shall use the following notation throughout the rest of the paper. For a topological group G with K a closed subgroup and X a nice topological space, we put

$$\begin{aligned}
\mathrm{Aut}(G) &= \text{Continuous automorphisms of } G\\
\mathrm{Inn}(G) &= \text{Inner automorphisms of } G\\
\mathrm{Out}(G) &= \mathrm{Aut}(G)/\mathrm{Inn}(G)\\
N_G(K) &= \text{Normalizer of } K \text{ in } G\\
C_G(K) &= \text{Centralizer of } K \text{ in } G\\
\mathrm{Aut}(G,K) &= \{\alpha \in \mathrm{Aut}(G)\colon \alpha|_K \in \mathrm{Aut}(K)\}\\
\mathrm{Inn}(G,K) &= \mathrm{Inn}(G) \cap \mathrm{Aut}(G,K)\\
\mathrm{Out}(G,K) &= \mathrm{Aut}(G,K)/\mathrm{Inn}(G,K)\\
\mu(a) &= \text{Conjugation by } a\text{; so, } \mu(a)(x) = axa^{-1} \text{ for } x \in G\\
\mathrm{TOP}(X) &= \text{The group of self-homeomorphisms of } X\\
\mathrm{Diff}(X) &= \text{The group of self-diffeomorphisms of a smooth manifold } X
\end{aligned}$$

2. Definitions, motivation and examples

2.1. *Rough definition of Seifert fiber spaces*

In Section 2, we explain what we mean by Seifert fiber spaces, give motivations, precise definitions and some examples.

The Seifert fiber space construction was first defined and studied in [12] and [13] and later reformulated in [34,38] and [41]. It is really a generalization of the classical Seifert spaces and homogeneous spaces of a Lie group by its uniform lattices. We shall start with a rough definition, continue with some motivation, and finally, give an exact definition.

We then examine what one must do to effectively construct all possible injective Seifert fiberings over a given base with a given typical fiber. As a result of this investigation, we determine the precise nature of the singular fibers.

A Seifert fibering $F \to E \xrightarrow{p} B$ is a "fibering" with singularities. For each $b \in B$, there is associated a finite subgroup Q_b of TOP(F), the group of self-homeomorphisms of F, so that $p^{-1}(b) = F/Q_b$. If Q_b is trivial, $p^{-1}(b) = F$ is called a typical fiber. Otherwise F/Q_b is called a singular fiber. For example, if $F = T^k$ is a k-torus, and Q_b acts freely on F, then F/Q_b will be a flat manifold which is finitely covered by the torus. We shall require the groups Q_b to be controlled in a particular way as a replacement of the local triviality condition. Before giving a formal definition, we start with some examples.

EXAMPLE 2.1.1. As a simple example, consider the Möbius band

$$E = I \times [-1, 1]/(0, t) \sim (1, -t).$$

It has an obvious circle action so that its orbit space $B = S^1 \setminus E$ is an interval. Thus, we get $S^1 \to E \xrightarrow{p} B$. The orbits are circles parallel to the boundary circle. The center circle orbit however is doubly covered by each of the other orbits. B is an arc $[0, 1]$. Here $p^{-1}(t) = S^1$, a free orbit for every $t \neq 0$, while $p^{-1}(0) = \mathbb{Z}_2 \setminus S^1$, the center circle, is the only singular fiber.

EXAMPLE 2.1.2. Let $\alpha : T^2 \to T^2$ be the map $\alpha(z_1, z_2) = (\bar{z}_1, \bar{z}_2)$ on a 2-torus. Then α has period 2, and has 4 fixed points. In fact, $B = T^2/\mathbb{Z}_2$ is a sphere with 4 branch points. Take the product $T^2 \times I$, a 2-torus and the unit interval. Identify the two ends by

$$(w, 0) \leftrightarrow \big(\alpha(w), 1\big)$$

to form the mapping torus M of α. Denote an element of M by $[w, t]$, and define an S^1-action on M by

$$\mathrm{e}^{2\pi i\theta}\big([w, t]\big) = [w, t + 2\theta],$$

where the second coordinate is taken modulo 1 and compatible with the identifications. The space $E = (T^2 \times I)/\mathbb{Z}_2$ is homeomorphic to an orientable 3-dimensional flat Riemannian manifold with holonomy group $\mathbb{Z}_2$. The circle S^1 acts as isometries on E. If p denotes the orbit mapping, then the orbit space is $B = T^2/\mathbb{Z}_2$. The free orbits, away from the 4 branch points, are typical fibers while the doubly covered orbits over the branch branch points are singular fibers. These singular fibers are still circles, but are half of the length of the typical fibers.

2.2. *Torus actions*

To facilitate our understanding of Seifert fiber spaces we shall examine first the situation arising from group actions. We shall assume familiarity with elementary properties of group actions. Good references as we shall use these concepts are the book by Bredon [7], the paper [12], and the chapter of this Handbook by Adem and Davis [1]. We recall first several definitions and facts.

Our spaces X are path-connected, completely regular and Hausdorff so that the various slice theorems are valid. For covering space theory, we need and tacitly assume that our spaces are locally path-connected and semi 1-connected.

A topological group G acts *properly* on X if the closure of $\{g \in G \mid gC \cap C \neq \emptyset\}$ is a compact subset of G, for all compact subsets C of X. See [49] for properties of proper actions. In particular, a proper action of a (not necessarily connected) locally compact Lie group on a completely regular space admits a slice, and the orbit space is Hausdorff.

If Q is a discrete group acting properly on X, we shall say that Q acts *properly discontinuously* on X. Note the isotropy subgroups are finite for a properly discontinuous action.

Let G be a path-connected group acting on a path-connected Hausdorff space X. Fix $x \in X$. The evaluation map $\mathrm{ev}^x : (G, e) \to (X, x)$ is defined by

$$\mathrm{ev}^x(g) = g \cdot x.$$

This induces the evaluation homomorphism

$$\mathrm{ev}^x_* : \pi_1(G, e) \to \pi_1(X, x).$$

Put H = image of $\mathrm{ev}^x_* \subset \pi_1(X, x)$. Then H is a central subgroup of $\pi_1(X, x)$ which is independent of choices of x [12, Lemma 4.2].

Let K be a normal subgroup of $\pi_1(X, x)$ and X_K the covering space of X associated with the subgroup K so that $\pi_1(X_K, x') = K$. Assume that $H \subset K$ and put

$$Q = \pi_1(X, x)/\pi_1(X_K, x') = \pi_1(X, x)/K.$$

Then we have the

PROPOSITION 2.2.1 ([12, §4] and [7, Chapter I, §9]). (1) *The G-action on X lifts to an action of X_K which commutes with the covering Q-action on X_K. The lifted action of G on X_K is proper if the action of G on X is proper.*

(2) *The induced Q action on the quotient space $W = G \setminus X_K$ is properly discontinuous (i.e., behaves at least like branched coverings with Hausdorff quotient) provided G acts properly on X.*

DEFINITION 2.2.2. A torus action (T^k, X) is said to be *injective* if $\mathrm{ev}^x_* : \pi_1(T^k, e) \to \pi_1(X, x)$ is injective.

Suppose T^k acts injectively on a path connected, locally path connected, semi 1-connected paracompact Hausdorff space X. Then, by the above proposition, the T^k action lifts to an action on $X_{\mathbb{Z}^k}$, where $\mathbb{Z}^k$ is the image of $\pi_1(T^k) \to \pi_1(M)$. We show that the lifted action $(T^k, X_{\mathbb{Z}^k})$ is free and splits.

PROPOSITION 2.2.3 ([12, §3.1]). *If T^k acts injectively on X, then $X_{\mathbb{Z}^k}$ splits into $T^k \times W$ so that $(T^k, X_{\mathbb{Z}^k}) = (T^k, T^k \times W)$, where the T^k action on $T^k \times W$ is via translation on the first factor and trivial on the simply connected W factor.*

PROOF. Let $G = T^k$ and S^1 be a subgroup of G isomorphic to the circle. (Note that S^1 is a direct factor.) Lift the S^1 action to $X' = X_{\mathbb{Z}^k}$. Let $y' \in X'$ and suppose $S^1_{y'} \neq 1$ is the stabilizer of the lifted S^1 action at y'. Choose paths $\tilde{\gamma}$ in X' from the base point x' over x to y' and $\alpha : (I, 0, 1) \to (S^1, 1, z)$ where z is the "first" element $\neq 1$ of S^1 for which $z \cdot y' = y'$. Then $\alpha(t)y'$ defines a loop at y'; and $\gamma * \alpha * \bar{\gamma}$ is the associated loop based at x'. Now $(\gamma * \alpha * \bar{\gamma})^n \sim \gamma * \alpha^n * \bar{\gamma}$ represents a generator of $\mathbb{Z}^k = H$. Hence n is the order of z in S^1. This implies that $\gamma * \alpha * \bar{\gamma}$ represents an nth root of a generator of $\mathbb{Z}^k$ which is impossible unless $n = 1$. Thus $S^1_{y'} = 1$, for all $y' \in X'$ and all circle subgroups of S^1 of T^k. Hence the lifted toral action (T^k, X') must be a free action and $X \to W = T^k \setminus X'$ is a principal T^k bundle.

We now show that this bundle is trivial. Let $f : W \to B_{T^k}$ be the classifying map from W into the classifying space for principal T^k-bundles. The bundle $X' \to W$ is represented by the homotopy class $[f] \in [W, B_{T^k}]$. Since B_{T^k} is a $K(\mathbb{Z}^k, 2)$ space, $[f]$ is represented by the image $f^*(u)$, where $u \in H^2(B_{T^k}; \mathbb{Z}^k)$ represents $\mathrm{id} : B_{T^k} \to B_{T^k}$.

The cohomology sequence,

$$\begin{aligned} 0 \longrightarrow H^1(W; \mathbb{Z}^k) \longrightarrow H^1(X'; \mathbb{Z}^k) \longrightarrow H^1(T^k; \mathbb{Z}^k) \longrightarrow H^2(W; \mathbb{Z}^k) \\ \longrightarrow H^2(X'; \mathbb{Z}^k), \end{aligned}$$

which arises from terms of low degree of the spectral sequence of the fibering $X' \to W$, is exact (e.g., [43, p. 332, Ex. 2], or [11, p. 329 C]).

We observe that $H^1(X'; \mathbb{Z}^k) \to H^1(T^k; \mathbb{Z}^k)$ is a surjective isomorphism because $\pi_1(T^k, 1) \to \pi_1(X', x')$ is an isomorphism. Consequently, $f^*(u)$ injects into $H^2(X'; \mathbb{Z}^k)$. On the other hand, $\pi^*(f^*(u)) = \tilde{f}^*(\pi'^*(u))$, where $\pi' : E_{T^k} \to B_{T^k}$, the universal T^k-bundle, and $\tilde{f}$ is the bundle map $X' \to E_{T^k}$ induced by f. But $\pi'^*(u) \in H^2(E_{T^k}; \mathbb{Z}^k) = 0$. Therefore, $\pi^*(f^*(u)) = 0$, which implies $f^*(u) = 0$. Hence, $X' \to W$ is a trivial fibering and $X' = T^k \times W$.

Finally, we remark that the group $T^k \times Q$ is acting properly on $T^k \times W$, and W is simply connected. The T^k action does not lift to the universal covering $\widetilde{X}$ of X but the induced ineffective $\mathbb{R}^k$ action on $T^k \times W$ lifts to an effective $\mathbb{R}^k$ action on $\mathbb{R}^k \times W$ and commutes with the group $\pi_1(X, x)$ of covering transformations on $\mathbb{R}^k \times W$. □

2.3. *Examples of injective actions*

A manifold M^n is called *hyper-aspherical* [15] if there is a degree 1-map from M to a closed aspherical manifold of dimension n.

PROPOSITION 2.3.1. *Any effective torus action on the following closed connected manifolds is injective*:

(1) *All aspherical manifolds* [12].
(2) *All hyper-aspherical manifolds* [15].
(3) *More generally, all manifolds M^n with $\xi^* : H^n(K(\pi_1(M), 1); \mathbb{Q}) \to H^n(M; \mathbb{Q})$ surjective, where ξ^* is induced by the classifying map $\xi : M \to K(\pi_1(M), 1)$, with $\pi_1(M)$ torsion free. (The torsion freeness assumption of $\pi_1(M)$ can be dropped if we assume the action is smooth, cf.* [8] *and* [40].) *See* [8,15,20,28,59,62] *and* [40] *for such generalizations.*
(4) *Homologically Kähler manifolds all of whose isotropy subgroups are finite (e.g., holomorphic actions)* [6, p. 170], [13, p. 186].

If every circle action on a closed manifold M is injective, the only compact connected Lie groups that can act on M are tori. See [12] for a proof of this fact. Therefore tori are the only compact connected Lie groups that act effectively on those manifolds listed in (1), (2) and (3). Moreover, in each of these cases, $\pi_1(T^k)$ must inject into the center of $\pi_1(M)$ if T^k acts effectively. Consequently, the only compact Lie groups that can act effectively on these M for which the center of $\pi_1(M)$ is finite are finite groups.

2.4. *Injective Seifert fibering*

Perhaps the most important feature of an injective torus action (T^k, X) is the splitting $(T^k, T^k \times W)$ in Proposition 2.2.3. The universal covering group $\mathbb{R}^k$ acts ineffectively via $\mathbb{R}^k \to T^k$ on $T^k \times W$ and lifts to the action of $\mathbb{R}^k$ as left translations on $\mathbb{R}^k \times W$, the universal covering space of $T^k \times W$ and of X. This $\mathbb{R}^k$ action contains the $\mathbb{Z}^k$ covering transformations over $T^k \times W$ and commutes with the group $\pi = \pi_1(X)$ of covering transformations on X. Together $\mathbb{R}^k$ and π generate a subgroup $\mathbb{R}^k \cdot \pi$ of TOP($\mathbb{R}^k \times W$) isomorphic to $\mathbb{R}^k \times_{\mathbb{Z}^k} \pi$, since $\mathbb{Z}^k = \mathbb{R}^k \cap \pi$.

We have been describing these features of injective actions to motivate our definition of injective Seifert fiberings. It is apparent that we could start with $\mathbb{R}^k \times W$ and using the reverse procedure, reconstruct the injective toral actions on X. We will now describe this reverse procedure. We will also put it in a more general context where $\mathbb{R}^k$ is replaced by any Lie group G and W is any completely regular space admitting covering space theory.

Let π be a discrete subgroup of TOP($G \times W$) and assume:

1. The left translational action $\ell(G)$ of G on $G \times W$ is *normalized* by π.
2. $\Gamma \subset \pi \cap \ell(G)$ is discrete and normal in π.
3. The induced action of $Q = \pi/\Gamma$ on W is proper.

These conditions imply that the group $\ell(G) \cdot \pi \subset$ TOP($G \times W$), generated by G and π, acts properly on $G \times W$.

With the above three conditions, we obtain the commutative diagram:

$$\begin{array}{ccccccl} G & \longrightarrow & G \times W & \xrightarrow{G\backslash} & W & & \text{Product principal } G\text{-fibering} \\ \downarrow{\scriptstyle \Gamma\backslash} & & \downarrow{\scriptstyle \pi\backslash} & & \downarrow{\scriptstyle Q\backslash} & & \\ \Gamma \backslash G & \longrightarrow & X = \pi \backslash (G \times W) & \xrightarrow{p} & Q \backslash W & & \text{Seifert fibering} \end{array}$$

X is called an *injective Seifert fiber space*, with *typical fiber* $\Gamma \backslash G$, B is called the *base* and the mapping p is called the *injective Seifert fibering*. Since the actions of π on $G \times W$ and the action of Q on W are properly discontinuous, the quotient spaces X and B are reasonable spaces.

If π centralizes $\ell(G)$, then $\Gamma \subset \pi \cap \ell(G)$ is in the center of G and the G action descends to a G/Γ action on X. In general, neither Γ is central in G nor π centralizes the G action but, as we shall see, the obvious "fibers" (inverse images $p^{-1}(b)$) still have a very nice description.

2.5. $\mathrm{TOP}_G(G \times W)$

To describe or construct all possible injective Seifert fiberings over B with typical fiber $\Gamma \backslash G$, we begin with a proper action of a discrete group Q on W so that $B = Q \backslash W$. We form the product space $G \times W$ letting G act on the first factor of $G \times W$ by left translations. Denote this action by $\ell(G) \subset \mathrm{TOP}(G \times W)$. We then search for discrete group π contained in the normalizer of $\ell(G)$ in $\mathrm{TOP}(G \times W)$ such that $\Gamma \subset \pi \cap G$ is discrete in G and the induced action of the quotient π/Γ is equivalent to the Q action on W. Our first task then is to describe the algebraic structure of the normalizer of $\ell(G)$ in $\mathrm{TOP}(G \times W)$. The topological structure of $\mathrm{TOP}(G \times W)$, while important and significant for some applications, is not necessary for our problem now at hand.

A space is *admissible* if it is completely regular (Hausdorff) locally path-connected and semi-locally simply-connected. Let G be a Lie group, $\mathrm{Aut}(G)$ the Lie group of continuous automorphisms of G. $\mathrm{TOP}(X)$ denotes the group of homeomorphisms of X. See Section 1 for a list of notation.

Recall that a homeomorphism f of $G \times W$ onto itself is *weakly G-equivariant* if and only if there exists a continuous automorphism α_f of G so that

$$f(a \cdot x, w) = \alpha_f(a) f(x, w)$$

for all $a \in G$ and $(x, w) \in G \times W$.

DEFINITION 2.5.1. The group of all weakly G-equivariant self-homeomorphisms of $G \times W$ is denoted by $\mathrm{TOP}_G(G \times W)$.

LEMMA 2.5.2. *$\mathrm{TOP}_G(G \times W)$ is the normalizer of $\ell(G)$ in* $\mathrm{TOP}(G \times W)$.

PROOF. Let $f \in \mathrm{TOP}(G \times W)$. Then the following are all equivalent:

(1) $f \in \mathrm{TOP}_G(G \times W)$.
(2) f is weakly G-equivariant.
(3) There exists $\alpha_f \in \mathrm{Aut}(G)$ such that $f(a \cdot x, w) = \alpha_f(a) f(x, w)$ for all $a \in G$ and $(x, w) \in G \times W$.
(4) There exists $\alpha_f \in \mathrm{Aut}(G)$ such that $f \circ \ell(a) \circ f^{-1} = \ell(\alpha_f(a))$ for all $a \in G$.
(5) f normalizes $\ell(G)$. □

Each element $f \in \mathrm{TOP}_G(G \times W)$ sends fibers of $G \times W \to W$ to fibers. That is, each f induces a map $\bar{f} \in \mathrm{TOP}(W)$ so that

$$\begin{array}{ccc} G \times W & \xrightarrow{f} & G \times W \\ \downarrow & & \downarrow \\ W & \xrightarrow{\bar{f}} & W \end{array}$$

commutes. Therefore, to each $f \in \mathrm{TOP}_G(G \times W)$, we can assign a pair $(\alpha_f, \bar{f}) \in \mathrm{Aut}(G) \times \mathrm{TOP}(W)$. This assignment

$$\mathrm{TOP}_G(G \times W) \longrightarrow \mathrm{Aut}(G) \times \mathrm{TOP}(W)$$

is a surjective homomorphism and it splits. For, if we define $(\alpha, h)(x, w) = (\alpha(x), h(w))$ for $(\alpha, h) \in \mathrm{Aut}(G) \times \mathrm{TOP}(W)$, then

$$\begin{aligned} (\alpha, h)(a \cdot x, w) &= \big(\alpha(a \cdot x), h(w)\big) \\ &= \big(\alpha(a)(\alpha(x), h(w)\big) \\ &= \ell\big(\alpha(a)\big)\big(\alpha(x), h(w)\big) \\ &= \ell\big(\alpha(a)\big)(\alpha, h)(x, w). \end{aligned}$$

Let K be the kernel of $\mathrm{TOP}_G(G \times W) \to \mathrm{Aut}(G) \times \mathrm{TOP}(W)$. If $f \in K$, since f moves only along the fibers, there exists a unique continuous function $\lambda : G \times W \to G$ such that $f(x, w) = (x \cdot (\lambda(x, w))^{-1}, w)$. We show that λ only depends upon W. For any $a \in G$,

$$\begin{aligned} ax \cdot \big(\lambda(ax, w)^{-1}, w\big) &= f(a \cdot x, w) \\ &= \ell(a) f(x, w) \quad \text{since } \alpha_f = \mathrm{id} \\ &= \ell(a)\big(x \cdot \big(\lambda(x, w)^{-1}, w\big)\big) \\ &= ax \cdot \big(\lambda(x, w)^{-1}, w\big). \end{aligned}$$

So, $\lambda(ax, w) = \lambda(x, w)$ which means λ only depends upon W. Let

$$\mathrm{M}(W, G) = \{\text{the continuous maps of } W \text{ into } G\}.$$

For each $\lambda \in \mathrm{M}(W, G)$, define $f_\lambda \in K$ by $f_\lambda(x, w) = (x \cdot (\lambda(w))^{-1}, w)$. The assignment $\mathrm{M}(W, G) \ni \lambda \mapsto f_\lambda \in K$ is easily checked to be an isomorphism, where the group law in $\mathrm{M}(W, G)$ is

$$(\lambda_1 \cdot \lambda_2)(w) = \lambda_1(w) \cdot \lambda_2(w).$$

Conjugation in $\mathrm{TOP}_G(G \times W)$ induces the action of $\mathrm{Aut}(G) \times \mathrm{TOP}(W)$ on $\mathrm{M}(W, G)$. It is given by

$$^{(\alpha,h)}\lambda = \alpha \circ \lambda \circ h^{-1},$$

that is,

$$(\alpha, h) \circ \lambda \circ (\alpha, h)^{-1}(x, w) = \big(x \cdot \alpha\big(\lambda\big(h^{-1}(w)\big)\big)^{-1}, w\big),$$

for

$$\begin{aligned}(\alpha, h) \circ \lambda \circ (\alpha, h)^{-1}(x, w) &= (\alpha, h) \circ \lambda\big(\alpha^{-1}(x), h^{-1}(w)\big) \\ &= (\alpha, h)\big(\alpha^{-1}(x) \cdot \big(\lambda\big(h^{-1}(w)\big)\big)^{-1}, h^{-1}(w)\big) \\ &= \big(x \cdot \alpha\big(\lambda\big(h^{-1}(w)\big)\big)^{-1}, w\big)\end{aligned}$$

and

$$\begin{aligned}\big(^{(\alpha,h)}\lambda\big)(x, w) &= \alpha \circ \lambda \circ h^{-1}(x, w) \\ &= \big(x \cdot \alpha\big(\lambda\big(h^{-1}(w)\big)\big)^{-1}, w\big).\end{aligned}$$

Therefore we have shown

LEMMA 2.5.3. $\mathrm{TOP}_G(G \times W) = \mathrm{M}(W, G) \rtimes (\mathrm{Aut}(G) \times \mathrm{TOP}(W))$, *and it is the normalizer of* $\ell(G)$ *in* $\mathrm{TOP}(G \times W)$.

The group operation in $\mathrm{TOP}_G(G \times W)$ is

$$\begin{aligned}(\lambda_1, \alpha_1, h_1) \cdot (\lambda_2, \alpha_2, h_2) &= \big(\lambda_1 \cdot {}^{(\alpha_1, h_1)}\lambda_2, \alpha_1 \circ \alpha_2, h_1 \circ h_2\big) \\ &= \big(\lambda_1 \cdot \big(\alpha_1 \circ \lambda_2 \circ h_1^{-1}\big), \alpha_1 \circ \alpha_2, h_1 \circ h_2\big).\end{aligned}$$

Specifically, (λ, α, h) acts on (x, w) by

$$\begin{aligned}(\lambda, \alpha, h) \cdot (x, w) &= \big((\lambda, 1, 1) \circ (1, \alpha, h)\big)(x, w) \\ &= (\lambda, 1, 1)\big(\alpha(x), h(w)\big) \\ &= \big(\alpha(x) \cdot \big(\lambda\big(h(w)\big)\big)^{-1}, h(w)\big).\end{aligned}$$

Then $\mathrm{M}(W, G) \rtimes \mathrm{Aut}(G)$ is the group of all weakly G-equivariant homeomorphisms of $G \times W$ which move only along the fibers. The group $\mathrm{M}(W, G)$ is the gauge group for the trivial fiber bundle $G \times W \to W$.

For $a \in G$, the constant map $W \to G$ sending W to a is denoted by $r(a)$. Clearly,

$$r(a) = (a, 1, 1) \in \mathrm{M}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{TOP}(W)\big).$$

This is a right translation by a^{-1} on the first factor of $G \times W$ so that $r(a)(x, w) = (x \cdot a^{-1}, w)$, and the subgroup of all such right translations is denoted by $r(G) \subset \mathrm{M}(W, G)$. Let $\ell(G)$ denote the group of left translations on the first factor so that $\ell(a)(x, w) = (a \cdot x, w)$. Then elements of $\ell(G)$ are of the form

$$\ell(a) = \big(a^{-1}, \mu(a), 1\big) \in \mathrm{M}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{TOP}(W)\big),$$

where $\mu(a) \in \mathrm{Inn}(G)$ is conjugation by a.

REMARK 2.5.4. If W is a smooth manifold and we take $\mathrm{TOP}_G(G \times W) \cap \mathrm{Diff}(G \times W)$, then this coincides with the weakly G-equivariant diffeomorphisms of $G \times W$ and

$$\mathrm{Diff}_G(G \times W) = \mathcal{C}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{Diff}(W)\big),$$

where $\mathcal{C}(W, G)$ is the group of smooth maps of W into G.

2.6. *Example: 3-dimensional Seifert manifolds with base the 2-torus*

We take $W = \mathbb{R}^2$. A group $Q = \mathbb{Z}^2$ acts on W as translations. Let

$$1 \to \mathbb{Z} \to \pi \to Q \to 1$$

be a central extension of $\mathbb{Z}$ by Q. Then π has a presentation

$$\pi = \langle \alpha, \beta, \gamma \mid [\alpha, \beta] = \gamma^p,\ [\alpha, \gamma] = [\beta, \gamma] = 1 \rangle,$$

where γ is a generator of the center $\mathbb{Z}$, and the images of α, β in Q are generators of Q. Suppose $p \neq 0$. Using $\mathbb{Z} \subset \mathbb{R}$, one can obtain an effective action of π on the product $\mathbb{R} \times W$ as follows: For $(z, x, y) \in \mathbb{R} \times W$,

$$\begin{aligned} \alpha(z, x, y) &= (z + y, x + 1, y), \\ \beta(z, x, y) &= (z, x, y + 1), \\ \gamma(z, x, y) &= \left(z + \frac{1}{p}, x, y\right). \end{aligned} \tag{2.1}$$

Notice that these maps are of the form

$$(z, x, y) \mapsto \big(\phi(z) - \lambda\big(h(x, y)\big), h(x, y)\big),$$

where ϕ is an automorphism of $\mathbb{R}$ and h is an action of Q on W, and λ is a map $W \to \mathbb{R}$. In our case ϕ is always the identity automorphism. Consequently, the group π lies in $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times W)$ as

$$(\lambda, \phi, h) \in \mathrm{M}(W, \mathbb{R}) \rtimes \big(\mathrm{GL}(1, \mathbb{R}) \times \mathrm{TOP}(W)\big).$$

The action of π on $\mathbb{R} \times W$ described above can be explained differently as follows. Consider the Heisenberg group

$$N = \left\{ \begin{bmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} : x, y, z \in \mathbb{R} \right\}$$

which is connected, simply connected and two-step nilpotent. We denote such a matrix by (z, x, y) so that

$$\begin{bmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} \quad \longleftrightarrow \quad (z, x, y).$$

Then the group operation is

$$(z', x', y') \cdot (z, x, y) = (z' + z + x'y,\ x' + x,\ y' + y),$$

and the center of N is 1-dimensional $\mathcal{Z} = \mathbb{R}$, consisting of all matrices with $x = y = 0$. The quotient $W = N/\mathcal{Z}$ is isomorphic to $\mathbb{R}^2$ so that

$$1 \to \mathbb{R} \to N \to \mathbb{R}^2 = W \to 1$$

is an exact sequence of groups. As spaces, this is a smooth fibration which is also a product $N = \mathbb{R} \times W$. Suppose $p \neq 0$ in the presentation of π. Let

$$\alpha = (0, 1, 0),\ \beta = (0, 0, 1),\ \text{and}\ \gamma = (1/p, 0, 0) \in N.$$

Then these satisfy the relations of the group π so that π sits in N as a discrete subgroup, and furthermore, the action of π defined above is nothing but the left multiplication on N by elements of π.

Actually this means that $\pi \setminus N$ is a nilmanifold which is a matter that we shall explore more later. The free action of π commutes with the $\mathbb{R}$ action on $\mathbb{R} \times \mathbb{R}^2$ and $\mathbb{R} \cap \pi$ is the central $\mathbb{Z}$ subgroup generated by $\gamma = (\frac{1}{p}, 0, 0)$. The $\mathbb{R}$ action on $\mathbb{R} \times \mathbb{R}^2$ descends to an effective $S^1 = \langle \gamma \rangle \setminus \mathbb{R}$ action on $\pi \setminus (\mathbb{R} \times \mathbb{R}^2)$. The map $\pi \setminus (\mathbb{R} \times \mathbb{R}^2) \to \mathbb{Z}^2 \setminus \mathbb{R}^2 = T^2$ is the orbit mapping of a free S^1 action. It is the principal S^1-bundle over T^2 with Euler class $-p$.

2.7. *Injective Seifert constructions*

In Section 2.5, we described the structure of the group of weakly G-equivariant self homeomorphisms of $G \times W$ as

$$\mathrm{TOP}_G(G \times W) = \mathrm{M}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{TOP}(W)\big)$$

and showed that it was the normalizer of $\ell(G)$ in $\mathrm{TOP}(G \times W)$. The action of $\mathrm{M}(W, G) \rtimes (\mathrm{Aut}(G) \times \mathrm{TOP}(W))$ on $G \times W$ is given by

$$(\lambda, \alpha, h) \cdot (x, w) = \big(\alpha(x) \cdot \big(\lambda(hw)\big)^{-1}, hw\big).$$

We observed that $\ell(G)$ embeds in $\mathrm{TOP}_G(G \times W)$ as

$$\ell_a \mapsto \big(a^{-1}, \mu(a), 1\big) \in \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) \subset \mathrm{TOP}_G(G \times W),$$

where $\mathrm{Inn}(G)$ denotes the inner automorphisms of G. Both $\mathrm{Inn}(G)$ and $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ are normal in $\mathrm{M}(W, G) \rtimes \mathrm{Aut}(G)$ and in $\mathrm{TOP}_G(G \times W)$, respectively. We may therefore rewrite $\mathrm{TOP}_G(G \times W)$ as the group extension

$$1 \to \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) \to \mathrm{TOP}_G(G \times W) \to \mathrm{Out}(G) \times \mathrm{TOP}(W) \to 1$$

where $\mathrm{Out}(G) = \mathrm{Aut}(G)/\mathrm{Inn}(G)$.

In Section 2.4, we described the condition that must be satisfied if a discrete subgroup π of $\mathrm{TOP}_G(G \times W)$ is to yield an injective Seifert fibering.

Suppose:

1. G is a Lie group.
2. Γ is isomorphic to a discrete subgroup of G.
3. W is an admissible space (see Section 2.5).
4. Q is a discrete group acting properly on W via $\rho : Q \to \mathrm{TOP}(W)$.
5. $1 \to \Gamma \to \pi \to Q \to 1$ is an extension.

DEFINITION. A homomorphism of π into $\mathrm{TOP}_G(G \times W)$ so that the diagram of extensions

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi \times \rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

commutes is called an *injective Seifert Construction*. The group $\mathrm{TOP}_G(G \times W)$ is called the *universal group* for this Seifert Construction. It follows that the group $\ell(G) \cdot \theta(\pi)$, generated by $\ell(G)$ and $\theta(\pi)$ in $\mathrm{TOP}_G(G \times W)$, acts properly on $G \times W$. The construction then yields the injective Seifert fibering

$$\ell(\Gamma) \backslash \ell(G) \longrightarrow \theta(\pi) \backslash (G \times W) \stackrel{p}{\longrightarrow} Q \backslash W$$

with the typical fiber $\Gamma \backslash G$ and base $Q \backslash W$.

Note that the total space of our Seifert fibering is $\pi \setminus (G \times W) = Q \setminus ((\Gamma \setminus G) \times W)$. We say it is *modelled on* $G \times W$. The base space is denoted by $B = Q \setminus W$.

2.8. *The singular fibers*

PROPOSITION 2.8.1. *Each singular fiber is the quotient of the homogeneous space $\Gamma \setminus G$ by a finite group of affine diffeomorphisms.*

PROOF. There is an intermediate fibering $\Gamma \setminus G \to (\Gamma \setminus G) \times W \to W$ so that the following diagram is commutative:

$$\begin{array}{ccccll}
G & \longrightarrow & G \times W & \longrightarrow & W & \\
 & & \downarrow{\scriptstyle \Gamma\setminus} & & \downarrow{\scriptstyle =} & \\
\Gamma \setminus G & \longrightarrow & (\Gamma \setminus G) \times W & \longrightarrow & W & \text{fibering} \\
 & & \downarrow{\scriptstyle Q\setminus} & & \downarrow{\scriptstyle Q\setminus} & \\
\Gamma \setminus G & \longrightarrow & \pi \setminus (G \times W) & \longrightarrow & Q \setminus W = B & \text{Seifert fibering}
\end{array}$$

The typical fiber of the Seifert fibering

$$\Gamma \setminus G \longrightarrow E \xrightarrow{p} B$$

is the homogeneous space $\Gamma \setminus G$. Then, what are the singular fibers?

(1) Pick $w \in W$ over $b \in B$.

(2) Over w, find $(\Gamma \setminus G) \times w$ in $(\Gamma \setminus G) \times W$.

(3) Let Q_w be the stabilizer of w. Then $p^{-1}(b) = Q_w \setminus ((\Gamma \setminus G) \times w)$.

Another way:

(4) Over w, find $G \times w$ in $G \times W$.

(5) Let π_w be the subgroup of π leaving $G \times w$ invariant. That is, $1 \to \Gamma \to \pi_w \to Q_w \to 1$ is the pullback of $1 \to \Gamma \to \pi \to Q \to 1$ via $Q_w \hookrightarrow Q$.

Then $p^{-1}(b) = Q_w \setminus ((\Gamma \setminus G) \times w) = \pi_w \setminus (G \times w)$.

The diagram (of the Seifert fibering) restricted to $G \times w$ is

$$\begin{array}{ccccc}
G & \longrightarrow & G \times w & \longrightarrow & w \\
 & & \downarrow & & \downarrow{\scriptstyle =} \\
\Gamma \setminus G & \longrightarrow & (\Gamma \setminus G) \times w & \longrightarrow & w \\
 & & \downarrow & & \downarrow \\
\Gamma \setminus G & \longrightarrow & \dfrac{(\Gamma \setminus G) \times w}{Q_w} & \longrightarrow & b \\
 & & \| & & \\
 & & \pi_w \setminus (G \times w) & &
\end{array}$$

We need to understand the action of Q_w on $(\Gamma \setminus G) \times w$, or equivalently, the action of π_w on $G \times w$.

The typical fibers, which are homogeneous spaces $\Gamma \setminus G$, occur over $b \in B$ for which $Q_w = 1$; that is, where Q acts freely. For example, if Q acts effectively on W and W is a connected manifold, then the set of typical fibers will be open and dense in E. Each of singular fibers is a quotient of $\Gamma \setminus G$ by a finite group of affine diffeomorphisms of G. This is because when one restricts $G \times W$ to $G \times w$, $\mathrm{M}(w, G) \rtimes (\mathrm{Aut}(G) \times \mathrm{TOP}(w))$ becomes $r(G) \rtimes \mathrm{Aut}(G)$.

Our group Γ injects into $\ell(G)$ and π_w goes into $\mathrm{M}(w, G) \rtimes \mathrm{Aut}(G)$ which is the same as $r(G) \rtimes \mathrm{Aut}(G)$. But $r(G) \rtimes \mathrm{Aut}(G) = \ell(G) \rtimes \mathrm{Aut}(G) = \mathrm{Aff}(G)$, because $(a, \alpha) = (a, \mu(a)^{-1}) \circ (1, \mu(a) \circ \alpha)$ in $\mathrm{M}(W, G) \rtimes \mathrm{Aut}(G)$. Therefore, we have

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi_w & \longrightarrow & Q_w & \longrightarrow & 1 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 1 & \longrightarrow & \ell(G) & \longrightarrow & \mathrm{Aff}(G) & \longrightarrow & \mathrm{Aut}(G) & \longrightarrow & 1 \end{array}$$

Thus, π_w acts on $G \times w \subset G \times W$ as a group of affine diffeomorphisms of G. The singular fiber $p^{-1}(b) = \pi_w \setminus (G \times w)$ is an infra-homogeneous space (see Section 2.9) if the finite group $Q_w = \pi_w / \Gamma$ acts on $\Gamma \setminus G$ freely, and is an orbifold covered by $\Gamma \setminus G$ otherwise. □

2.9. *Infra-homogeneous spaces*

For a discrete $\pi \subset \mathrm{Aff}(G)$ for which π acts properly on G and $\pi \cap G = \Gamma$ is of finite index in π, we obtain an example of an injective Seifert fibering by choosing $W = \{w\}$, a point. We get

$$\pi \setminus \big(G \times \{w\}\big) \xrightarrow{p} B = \{w\}.$$

This is an injective Seifert fibering with only one fiber, which is singular, and with typical fiber the homogeneous $\Gamma \setminus (G \times \{w\})$.

We like to make more precise the geometry carried by the singular fibers of our Seifert fiberings. Endow G with the linear connection defined by the left-invariant vector fields. Since the parallel transport is the effect of the left translations on the tangent vectors of G, and hence clearly independent of paths; the connection is flat. A geodesic through the identity element $e \in G$ are the 1-parameter subgroups of G and thus defined for any real value of the affine parameter. All geodesics are translates of geodesics through e and thus the connection is complete. One easily checks that the torsion tensor has vanishing covariant derivative. According to [24, Proposition 2.1],

$$\mathrm{Aff}(G) = \ell(G) \rtimes \mathrm{Aut}(G)$$

is the group of affine diffeomorphisms (= connection-preserving diffeomorphisms) of G, and $(a, \alpha) \in G \rtimes \mathrm{Aut}(G)$ acts on G by $(a, \alpha)(x) = a \cdot \alpha(x)$ for all $x \in G$. For example, if $G = \mathbb{R}^n$, $\mathrm{Aff}(\mathbb{R}^n) = \mathbb{R}^n \rtimes \mathrm{GL}(n, \mathbb{R})$, the ordinary affine group of $\mathbb{R}^n$.

Suppose π is a discrete subgroup of $\mathrm{Aff}(G)$ and acts properly on G. Suppose $\Gamma \subset \pi \cap G$ and $Q = \pi/\Gamma$ is finite. If π acts freely on G, we call $\Gamma \backslash G$ a *homogeneous space* and the quotient $\pi \backslash G = Q \backslash (\Gamma \backslash G)$ an *infra-homogeneous space*. However in general, π will not necessarily act freely and $\pi \backslash G$ is then an orbifold (i.e., a V-manifold). One should notice that the finite regular (possibly branched) covering $\Gamma \backslash G \to \pi \backslash G$ is extremely nice since the action of Q came from $\mathrm{Aff}(G)$.

The case where G is a simply connected Abelian or nilpotent Lie group is especially important for us. For $G = \mathbb{R}^n$, $\mathrm{O}(n)$ is a maximal compact subgroup of $\mathrm{GL}(n, \mathbb{R})$. A uniform (= cocompact) discrete subgroup π of $\mathbb{R}^n \rtimes \mathrm{O}(n)$ is called a *crystallographic group*. By a theorem of Bieberbach, $\pi \cap \mathbb{R}^n$ is isomorphic to $\mathbb{Z}^n$, and is a lattice (= uniform discrete subgroup) of $\mathbb{R}^n$. [In general, a lattice Γ of a Lie group G is a discrete subgroup such that $\Gamma \backslash G$ has finite volume. For a solvable Lie group G, this is equivalent to $\Gamma \backslash G$ being compact.] If π is torsion-free, we call π a *Bieberbach group* (= torsion free crystallographic group). Flat manifolds are the orbit spaces $\pi \backslash \mathbb{R}^n$, where π is a Bieberbach group. Note that each flat manifold is finitely covered by a flat torus $\mathbb{Z}^n \backslash \mathbb{R}^n$.

Let G be a connected, simply connected nilpotent Lie group. Choose a maximal compact subgroup C of $\mathrm{Aut}(G)$. A uniform discrete subgroup π of $G \rtimes C$ is called an *almost crystallographic group*. If it is torsion-free, it is called an *almost Bieberbach group*. An almost Bieberbach group π yields an *infra-nilmanifold* $\pi \backslash G$. Note here again that any infra-nilmanifold is finitely covered by the nilmanifold $\Gamma \backslash G$, where $\Gamma = \pi \cap G$. See [39] and [2] for more details.

Thus, nilmanifolds are a generalization of tori, and infra-nilmanifolds correspond to flat manifolds. It is also known that a manifold is diffeomorphic to an infra-nilmanifold if and only if it is *almost flat*. This term is due to Gromov. See [17] for a proof of the above fact.

3. Group cohomology

3.1. *Introduction*

The injective Seifert Construction entails embedding a discrete group π into the universal group $\mathrm{TOP}_G(G \times W)$ in such a way that the following diagram of short exact sequences of groups commutes:

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi \times \rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

Therefore, in order to understand when such constructions are possible and to classify them when they do exist, we need to investigate the general procedure for mapping one short exact sequence of groups into another. These results will be formulated in terms of the first and second cohomology of groups with not necessarily Abelian coefficients. We will need the full generality of these results later as we build in more geometry to our Seifert Constructions by "reducing" our universal groups $\mathrm{TOP}_G(G \times W)$.

First, we shall recall some definitions and elementary properties of group extensions. Then we explain the correspondence between congruence classes of extensions of Abelian group A by a group Q in terms of the second cohomology of Q with coefficients in A, $H^2(Q; A)$. If A is replaced by a non-Abelian group G, we show that the congruence classes of extensions of G by Q, $\text{Opext}(G, Q, \varphi)$ is in one-one correspondence with $H^2(Q; \mathcal{Z}(G))$, where $\mathcal{Z}(G)$ is the center of G.

Next, we study the problem of mapping homomorphically one short exact sequence of groups into another. Given one such homomorphism θ_0, we show that all other possible homomorphisms are measured by the first cohomology group. Finally, if we change the extension of the initial exact sequence (measured by a second cohomology group) we determine when it too maps into the target exact sequence.

3.2. *Group extensions*

A group extension is a short exact sequence

$$1 \to G \to E \to Q \to 1$$

of not necessarily Abelian groups. There is a naturally associated homomorphism $\varphi : Q \to \text{Out}(G)$, called the *abstract kernel* of the extension. This comes about as follows. Pick a "section" $s : Q \to E$ so that the composite $Q \to E \to Q$ is the identity map (s is not necessarily a homomorphism). We pick $s(1) = 1$. This defines a map $\tilde{\varphi} : Q \to \text{Aut}(G)$ (which is a lift of the abstract kernel φ) given by

$$\tilde{\varphi}(\alpha) = \mu\big(s(\alpha)\big),$$

where μ is the conjugation map. Even if $\tilde{\varphi}$ is not a homomorphism, it induces a homomorphism $\varphi : Q \to \text{Out}(G)$, which is our abstract kernel. Of course, φ does not depend on the choice of the section s. We say *the group E is an extension associated with the abstract kernel $\varphi : Q \to \text{Out}(G)$.*

Define $f : Q \times Q \to G$ by

$$s(\alpha) \cdot s(\beta) = f(\alpha, \beta) \cdot s(\alpha\beta).$$

Then one can easily verify that

$$\tilde{\varphi}(\alpha) \circ \tilde{\varphi}(\beta) = \mu\big(f(\alpha, \beta)\big) \circ \tilde{\varphi}(\alpha\beta), \tag{3.1}$$

$$f(\alpha, 1) = 1 = f(1, \beta), \tag{3.2}$$

$$f(\alpha, \beta) \cdot f(\alpha\beta, \gamma) = \tilde{\varphi}(\alpha)\big(f(\beta, \gamma)\big) \cdot f(\alpha, \beta\gamma) \tag{3.3}$$

for every $\alpha, \beta, \gamma \in Q$. In fact, (3.1) follows from the definition of the map f above, (3.2) follows from $s(1) = 1$, and (3.3) is a result of the associative law of G.

Conversely, suppose we have maps $\tilde{\varphi} : Q \to \text{Aut}(G)$ (not necessarily a homomorphism) lifting a homomorphism $\varphi : Q \to \text{Out}(G)$ and $f : Q \times Q \to G$ satisfying (3.1), (3.2) and (3.3). Then there exists an extension $1 \to G \to E \to Q \to 1$ with the given abstract kernel φ. In fact, $E = G \times Q$ as a set, and has group operation

$$(a,\alpha)\cdot(b,\beta) = \big(a\cdot\tilde{\varphi}(\alpha)(b)\cdot f(\alpha,\beta), \alpha\beta\big).$$

Thus we have

PROPOSITION 3.2.1. *For a given abstract kernel* $\varphi: Q \to \mathrm{Out}(G)$, *pick a lift* $\tilde{\varphi}: Q \to \mathrm{Aut}(G)$. *Then the set of all extensions with abstract kernel* φ *is in one-one correspondence with the set of all maps* $f: Q \times Q \to G$ *satisfying* (3.1), (3.2) *and* (3.3).

DEFINITION 3.2.2. Let $\varphi: Q \to \mathrm{Out}(G)$ be a homomorphism, $\tilde{\varphi}: Q \to \mathrm{Aut}(G)$ be a lift, and $f: Q \times Q \to G$ a map satisfying (3.1), (3.2) and (3.3). Then the group $E = G \times Q$ with group operation

$$(a,\alpha)\cdot(b,\beta) = \big(a\cdot\tilde{\varphi}(\alpha)(b)\cdot f(\alpha,\beta), \alpha\beta\big).$$

is denoted by $G \times_{(f,\tilde{\varphi})} Q$.

Note that some abstract kernel $\varphi: Q \to \mathrm{Out}(G)$ may not have any extension at all. When there is a lift $\tilde{\varphi}$ which is a homomorphism (for example, when G is Abelian), there exists at least one extension (by taking $f = 1$, the constant map), the *semi-direct product* $G \times_{(1,\tilde{\varphi})} Q$. As a set, this is $G \times Q$, and the group operation is

$$(a,\alpha)\cdot(b,\beta) = \big(a\cdot\tilde{\varphi}(\alpha)(b), \alpha\beta\big).$$

DEFINITION 3.2.3. Two extensions

$$1 \to G \to E \to Q \to 1 \quad \text{and} \quad 1 \to G \to E' \to Q \to 1$$

with the same abstract kernels are *congruent* if there is an isomorphism $\theta: E \to E'$ which restricts to the identity map on G, and induces the identity on Q. That is, the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{=} & & \downarrow{\theta} & & \downarrow{=} & & \\ 1 & \longrightarrow & G & \longrightarrow & E' & \longrightarrow & Q & \longrightarrow & 1 \end{array}$$

is commutative.

DEFINITION 3.2.4. Let $\varphi: Q \to \mathrm{Out}(G)$ be a homomorphism, and $\tilde{\varphi}: Q \to \mathrm{Aut}(G)$ be a lift of φ. Then $\mathrm{Opext}(Q, G, \tilde{\varphi})$ denotes the set of all congruence classes of extensions of the group G by Q with operator $\tilde{\varphi}$.

Therefore, an element $[f] \in \mathrm{Opext}(Q, G, \tilde{\varphi})$ is represented by an extension $1 \to G \to E \to Q \to 1$ with $E = G \times_{(f,\tilde{\varphi})} Q$.

3.3. *Pullback and pushout*

3.3.1. (Pushout) Let Γ be a subgroup of G. We say (G, Γ) has UAEP (unique automorphism extension property) if every automorphism of Γ extends to an automorphism of G uniquely. For example, if G is a connected, simply connected nilpotent Lie group, and Γ is any lattice, then (G, Γ) has UAEP. In particular, $(\mathbb{R}^n, \mathbb{Z}^n)$, where $\mathbb{Z}^n$ spans $\mathbb{R}^n$, has UAEP.

Let $\varepsilon : \Gamma \hookrightarrow G$ be an injective homomorphism. Suppose $(G, \varepsilon(\Gamma))$ has UAEP, and

$$1 \to \Gamma \to \pi \to Q \to 1$$

is exact. Then there is a unique group E fitting the commuting diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \varepsilon} & & \downarrow & & \downarrow{\scriptstyle =} & & \\ 1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \end{array} \qquad \textit{pushout} \text{ by } \varepsilon : \Gamma \to G$$

Such a group E can be constructed as follows. Suppose $[f] \in \mathrm{Opext}(Q, \Gamma, \tilde{\varphi})$ is represented by the extension $\pi = \Gamma \times_{(f,\tilde{\varphi})} Q$. That is, $\pi = \Gamma \times Q$ as a set, and has group operation

$$(a, \alpha) \cdot (b, \beta) = \big(a \cdot \tilde{\varphi}(\alpha)(b) \cdot f(\alpha, \beta), \alpha\beta\big)$$

for $a, b \in \Gamma$ and $\alpha, \beta \in Q$. Recall that $\tilde{\varphi}$ and f satisfy the equalities (3.1), (3.2) and (3.3) in Section 3.2. Since (G, Γ) has the UAEP, $\tilde{\varphi} : Q \to \mathrm{Aut}(\Gamma)$ can be viewed as $\tilde{\varphi} : Q \to \mathrm{Aut}(\Gamma) \to \mathrm{Aut}(G)$ and $f : Q \times Q \to \Gamma$ as $f : Q \times Q \to \Gamma \xrightarrow{\varepsilon} G$. With these new interpretations of $\tilde{\varphi}$ and f, the equalities (3.1), (3.2) and (3.3) still hold. Now we can define E by $E = G \times_{(f,\tilde{\varphi})} Q$. That is, $E = G \times Q$ as a set, and has group operation

$$(a, \alpha) \cdot (b, \beta) = \big(a \cdot \tilde{\varphi}(\alpha)(b) \cdot f(\alpha, \beta), \alpha\beta\big)$$

for $a, b \in G$ and $\alpha, \beta \in Q$.

When ε is an obvious inclusion $i : \Gamma \hookrightarrow G$, we denote such an extension E simply by $G\pi$ so that

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle i} & & \downarrow{\scriptstyle \cap} & & \downarrow{\scriptstyle =} & & \\ 1 & \longrightarrow & G & \longrightarrow & G\pi & \longrightarrow & Q & \longrightarrow & 1 \end{array} \qquad \textit{pushout} \text{ by } i : \Gamma \to G$$

3.3.2. (Pullback) Let

$$1 \to G \to \widetilde{P} \to P \to 1$$

be exact, and $\rho: Q \to P$ be a homomorphism. Then there is a unique group $\widetilde{Q}$ fitting the commuting diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & \widetilde{Q} & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow {\scriptstyle =} & & \downarrow & & \downarrow {\scriptstyle \rho} & & \\ 1 & \longrightarrow & G & \longrightarrow & \widetilde{P} & \longrightarrow & P & \longrightarrow & 1 \end{array} \qquad \textit{pullback} \text{ by } \rho: Q \to P$$

Such a group $\widetilde{Q}$ can be constructed as follows. Suppose $[f] \in \mathrm{Opext}(P, G, \tilde{\varphi})$ is represented by the extension $\widetilde{P} = G \times_{(f,\tilde{\varphi})} P$. That is, $\widetilde{P} = G \times P$ as a set, and has group operation

$$(a, \alpha) \cdot (b, \beta) = \big(a \cdot \tilde{\varphi}(\alpha)(b) \cdot f(\alpha, \beta), \alpha\beta\big),$$

where $\tilde{\varphi}: P \to \mathrm{Aut}(G)$ and $f: P \times P \to G$. Now, $\tilde{\varphi}: P \to \mathrm{Aut}(G)$ gives rise to $\tilde{\varphi}: Q \stackrel{\rho}{\to} P \to \mathrm{Aut}(G)$, and $f: P \times P \to G$ gives rise to $f: Q \times Q \stackrel{\rho\times\rho}{\to} P \times P \to G$. With these new interpretations of $\tilde{\varphi}$ and f, the equalities (3.1), (3.2) and (3.3) still hold. Now we can define $\widetilde{Q}$ by, $\widetilde{Q} = G \times_{(f,\tilde{\varphi})} Q$. That is, $\widetilde{Q} = G \times Q$ as a set, and has group operation

$$(a, \alpha) \cdot (b, \beta) = \big(a \cdot \tilde{\varphi}(\alpha)(b) \cdot f(\alpha, \beta), \alpha\beta\big)$$

for $a, b \in G$ and $\alpha, \beta \in Q$.

3.4. *Extensions with Abelian kernel A and $H^2(Q; A)$*

Consider the case where $G = A$ is Abelian in the previous section. In other words, we try to classify all group extensions of an Abelian group A by Q with abstract kernel $\varphi: Q \to \mathrm{Out}(A)$. Since A is Abelian, $\mathrm{Aut}(A) = \mathrm{Out}(A)$, and there is no need of having a lift $\tilde{\varphi}$. Moreover (3.1) holds automatically. A map $f: Q \times Q \to A$ satisfying the two equalities (3.2) and (3.3) is called a *factor set* (or *2-cocycle*) for the abstract kernel φ. The set of all factor sets is denoted by $Z^2_\varphi(Q; A)$ or simply, by $Z^2(Q; A)$.

Let $\lambda: Q \to A$ be any map such that $\lambda(1) = 0$. Define $\delta\lambda: Q \times Q \to A$ by

$$\delta\lambda(\alpha, \beta) = \lambda(\alpha) + \varphi(\alpha)\big(\lambda(\beta)\big) - \lambda(\alpha\beta). \tag{3.4}$$

It turns out that such a $\delta\lambda$ is a 2-cocycle and is called a *2-coboundary*. The set of all 2-coboundaries is denoted by $B^2_\varphi(Q; A)$. Clearly, $B^2_\varphi(Q; A)$ is a subgroup of $Z^2_\varphi(Q; A)$. Let $f_1, f_2: Q \times Q \to A$ be 2-cocycles. We say f_1 is *cohomologous* to f_2 if $f_1 - f_2 \in B^2_\varphi(Q; A)$; that is, if there is a map $\lambda: Q \to A$ such that $f_2 - f_1 = \delta\lambda$.

We define the second cohomology group as the quotient group

$$H^2_\varphi(Q; A) = Z^2_\varphi(Q; A)/B^2_\varphi(Q; A).$$

Since the abstract kernel $\varphi : Q \to \mathrm{Out}(A)$ lifts to a homomorphism $\varphi : Q \to \mathrm{Out}(A) = \mathrm{Aut}(A)$, there is always an extension; namely, the semi-direct product $A \rtimes Q$. Then every 2-cocycle $f \in Z^2_{\varphi}(Q; A)$ gives rise to an extension of A by Q, which can be denoted by $A \times_{(f,\varphi)} Q$. Furthermore, it is easy to see that two cocycles f_1 and f_2 yield congruent extensions $A \times_{(f_1,\varphi)} Q$ and $A \times_{(f_2,\varphi)} Q$ if and only if f_1 and f_2 are cohomologous. Consequently, we have

$$\mathrm{Opext}(A, Q, \varphi) \equiv H^2_{\varphi}(Q; A)$$

as sets, and they classify the extensions of A by Q with abstract kernel φ, up to congruence. It is customary to identify the zero element of $H^2_{\varphi}(Q; A)$ with the extension class of the semi-direct product $A \rtimes Q$. Then addition of the 2-cocycles in $Z^2_{\varphi}(Q; A)$ induces the group operations in $H^2_{\varphi}(Q; A)$.

3.5. *Extensions with non-Abelian kernel G and $H^2(Q; \mathcal{Z}(G))$*

For any homomorphism $\varphi : Q \to \mathrm{Out}(G)$, take any lift $\tilde{\varphi} : Q \to \mathrm{Aut}(G)$. This, in turn, induces a unique homomorphism $\bar{\varphi} : Q \to \mathrm{Aut}(\mathcal{Z}(G))$ by restriction (regardless which lift of φ is chosen, and even if $\tilde{\varphi}$ was not a homomorphism). We shall define an action of $H^2_{\bar{\varphi}}(Q; \mathcal{Z}(G))$ on the set $\mathrm{Opext}(G, Q; \bar{\varphi})$ which will turn out to be simply transitive.

Let $g \in Z^2_{\bar{\varphi}}(Q; \mathcal{Z}(G))$; that is, $g : Q \times Q \to \mathcal{Z}(G)$ satisfying the equalities (3.2) and (3.3) of Section 3.2. For any $[f_0] \in \mathrm{Opext}(G, Q; \tilde{\varphi})$, define

$$(g \cdot f_0)(\alpha, \beta) = g(\alpha, \beta) \cdot f_0(\alpha, \beta).$$

Then, clearly, $g \cdot f_0$ satisfies the equalities (3.1), (3.2) and (3.3). Moreover, if two extensions $G \times_{(f_0,\tilde{\varphi})} Q$ and $G \times_{(f_1,\tilde{\varphi})} Q$ are congruent, so are $G \times_{(g \cdot f_0,\tilde{\varphi})} Q$ and $G \times_{(g \cdot f_1,\tilde{\varphi})} Q$. Therefore

$$(g, f_0) \mapsto g \cdot f_0$$

is an action of $Z^2(Q; \mathcal{Z}(G))$ on $\mathrm{Opext}(G, Q; \tilde{\varphi})$.

Suppose $[g \cdot f] = [f]$ for some $[f] \in \mathrm{Opext}(G, Q; \tilde{\varphi})$. Then one easily sees that $g \in B^2_{\bar{\varphi}}(Q; \mathcal{Z}(G))$. This gives us a simple action of $H^2_{\bar{\varphi}}(Q; \mathcal{Z}(G))$ on $\mathrm{Opext}(G, Q; \tilde{\varphi})$.

Now we show that this action is transitive. Suppose $f, f' \in \mathrm{Opext}(G, Q; \tilde{\varphi})$. Let

$$g(\alpha, \beta) = f(\alpha, \beta) \cdot f'(\alpha, \beta)^{-1}$$

for all $\alpha, \beta \in Q$. Then $g(\alpha, \beta)$ lies in $\mathcal{Z}(G)$, and it is easy to check

$$g : Q \times Q \to \mathcal{Z}(G)$$

satisfies the equalities (3.2) and (3.3). Thus, $g \in Z^2_{\tilde{\varphi}}(Q; \mathcal{Z}(G))$. Consequently, we have shown the action of $H^2_{\tilde{\varphi}}(Q; \mathcal{Z}(G))$ on $\mathrm{Opext}(G, Q; \tilde{\varphi})$ is simply transitive. Thus,

$$\mathrm{Opext}(G, Q, \tilde{\varphi}) \approx H^2_{\tilde{\varphi}}(Q; \mathcal{Z}(G))$$

as sets, and they classify the extensions of G by Q with abstract kernel φ, up to congruence. In general, however, there may not exist any extension of G by Q with the given abstract kernel. Moreover, even when there exists one, there is no *a priori* special element, like the semi-direct product $G \rtimes Q$, because the semi-direct product can be formed when and only when the abstract kernel lifts to an homomorphism into $\mathrm{Aut}(G)$.

3.6. $H^1(Q; C)$ *with a non-Abelian* C

The first cohomology is useful in classifying homomorphisms from one group to another. Let C be a group (not necessarily Abelian) and $\varphi : Q \to \mathrm{Aut}(C)$ be a homomorphism. We shall define the first cohomology set $H^1_{\varphi}(Q; C)$. A map $\eta : Q \to C$ is called a *1-cocycle* if it satisfies

$$\eta(\alpha\beta) = \eta(\alpha) \cdot \varphi(\alpha)(\beta)$$

for all $\alpha, \beta \in Q$. The set of all 1-cocycles is denoted by $Z^1_{\varphi}(Q; C)$. Two 1-cocycles $\eta, \eta' : Q \to C$ are *cohomologous* if there exists $c \in C$ such that

$$\eta'(\alpha) = c \cdot \eta(\alpha) \cdot \varphi(\alpha)\big(c^{-1}\big)$$

for all $\alpha \in Q$. The set of cohomologous classes of $Z^1_{\varphi}(Q; C)$ is denoted by $H^1_{\varphi}(Q; C)$. When C is Abelian, $H^1_{\varphi}(Q; C)$ is the ordinary first cohomology group.

3.7. *Mapping one group extension into another*

DEFINITION 3.7.1. A diagram of short exact sequences of groups

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\ & & \big\downarrow{\scriptstyle i} & & \big\downarrow{\scriptstyle \psi} & & \big\downarrow{\scriptstyle \bar{\theta}} & & \\ 1 & \longrightarrow & G' & \longrightarrow & E' & \longrightarrow & Q' & \longrightarrow & 1 \end{array} \tag{3.5}$$

is called *half-commutative* if i is an inclusion and

(A) the right-hand side square is commutative, and

(B) for every $x \in E$, $i(x \cdot a \cdot x^{-1}) = \psi(x) \cdot i(a) \cdot \psi(x)^{-1}$ for all $a \in G$.

Condition (B) is weaker than the left-hand side square being commutative (i.e., $\psi|_G = i$). It simply means that

$$i \circ \mu(x) = \mu'(\psi(x)) \circ i : G \to G'$$

for all $x \in E$. It is not required that $\psi|_G = i$.

Let $[f] \in \mathrm{Opext}(Q, G, \tilde{\varphi})$ represent the congruence class of the extension E. That is, $E = G \times_{(f,\tilde{\varphi})} Q$. The homomorphism $\psi : E \to E'$ gives rise to a map

$$\tilde{\varphi}' : Q \to \mathrm{Aut}(G')$$

by $\tilde{\varphi}'(\alpha) = \mu'(\psi(1,\alpha))$, where μ' is conjugation in E'. Since $\tilde{\varphi}(\alpha) = \mu(1,\alpha)$, the two maps $\tilde{\varphi} : Q \to \mathrm{Aut}(G)$ and $\tilde{\varphi}' : Q \to \mathrm{Aut}(G')$ are related by the commuting diagram

$$\begin{array}{ccc} G & \xrightarrow{\tilde{\varphi}(\alpha)} & G \\ \downarrow{\scriptstyle i} & & \downarrow{\scriptstyle i} \\ G' & \xrightarrow{\tilde{\varphi}'(\alpha)} & G' \end{array}$$

This shows that $\tilde{\varphi}'(\alpha)$ maps $i(G)$ onto itself, and hence $C = C_{G'}(i(G))$, the centralizer of $i(G)$ in G', onto itself also.

THEOREM 3.7.2. *Let $\psi : E \to E'$ be a homomorphism which makes the diagram* (3.5) *half commutative. Let $E = G \times_{(f,\tilde{\varphi})} Q$, and define $\tilde{\varphi}'(\alpha) = \mu'(\psi(1,\alpha))$ for $\alpha \in Q$ (μ'* is conjugation in *E'). Then,*

(1) *There exists a homomorphism $\theta : E \to E'$ making the diagram*

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle i} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \bar{\theta}} & & \\ 1 & \longrightarrow & G' & \longrightarrow & E' & \longrightarrow & Q' & \longrightarrow & 1 \end{array} \qquad (3.6)$$

commutative if and only if there exists a map $\lambda : Q \to C = C_{G'}(i(G))$ satisfying

$$i(f(\alpha,\beta),1) \cdot \lambda(\alpha\beta)^{-1} = \lambda(\alpha)^{-1} \cdot \tilde{\varphi}'(\alpha)(\lambda(\beta)^{-1}) \cdot \psi(f(\alpha,\beta),1) \qquad (3.7)$$

for all $\alpha, \beta \in Q$.

(2) *Suppose that there exists a homomorphism $\theta_0 : E \to E'$ as above. Then $\mu' \circ \theta_0$ induces a homomorphism $\varphi'' : Q \to$* Aut*(C).*

The set of all θ's fitting the diagram, up to conjugation by elements of G' (in fact, of C), is in one-one correspondence with $H^1_{\varphi''}(Q; C)$.

(3) *Suppose that there exists a homomorphism $\theta_0 : E \to E'$ as above, and that $i(\mathcal{Z}(G)) \subset \mathcal{Z}(G')$. Let $[E''] \in$* Opext*$(Q, G, \tilde{\varphi})$ be an extension. Then there exists a homomorphism $E'' \to E'$ completing the diagram if and only if*

$$[E''] - [E] \in \mathrm{Ker}\{i_* : H^2(Q; \mathcal{Z}(G)) \to H^2(Q; \mathcal{Z}(G'))\}.$$

PROOF. (1) The extension E is $G \times Q$ with group law

$$(a,\alpha)\cdot(b,\beta) = \big(a \cdot \tilde{\varphi}(\alpha)(b) \cdot f(\alpha,\beta), \alpha\beta\big).$$

Then

$$(1,\alpha)(a,1)(1,\alpha)^{-1} = \big(\tilde{\varphi}(\alpha)(a), 1\big).$$

Suppose there is a homomorphism $\theta : E \to E'$ making both squares commutative. Then

$$\begin{aligned}\theta(1,\alpha)\cdot i(a,1)\cdot\theta(1,\alpha)^{-1} &= \theta\big((1,\alpha)(a,1)(1,\alpha)^{-1}\big)\\ &= i\big((1,\alpha)(a,1)(1,\alpha)^{-1}\big)\\ &= \psi(1,\alpha)\cdot i(a,1)\cdot\psi(1,\alpha)^{-1}.\end{aligned}$$

Since conjugations of $i(a,1)$ by $\theta(1,\alpha)$ and $\psi(1,\alpha)$ are the same, their difference must lie in the centralizer of $i(G)$ in G'. In other words, there exists a map

$$\lambda : Q \to C = C_{G'}\big(i(G)\big)$$

so that $\theta(1,\alpha) = \lambda(\alpha)^{-1}\cdot\psi(1,\alpha)$. Thus, in general,

$$\begin{aligned}\theta(a,\alpha) &= \theta(a,1)\cdot\theta(1,\alpha)\\ &= i(a,1)\cdot\lambda(\alpha)^{-1}\cdot\psi(1,\alpha).\end{aligned}$$

Now

$$\begin{aligned}\theta\big((1,\alpha)(1,\beta)\big) &= \theta\big(f(\alpha,\beta),\alpha\beta\big)\\ &= i\big(f(\alpha,\beta),1\big)\cdot\lambda(\alpha\beta)^{-1}\cdot\psi(1,\alpha\beta).\end{aligned} \tag{3.8}$$

On the other hand,

$$\begin{aligned}\theta(1,\alpha)\cdot\theta(1,\beta) &= \big(\lambda(\alpha)^{-1}\cdot\psi(1,\alpha)\big)\cdot\big(\lambda(\beta)^{-1}\cdot\psi(1,\beta)\big)\\ &= \lambda(\alpha)^{-1}\cdot\big\{\psi(1,\alpha)\cdot\lambda(\beta)^{-1}\cdot\psi(1,\alpha)^{-1}\big\}\cdot\big\{\psi(1,\alpha)\cdot\psi(1,\beta)\big\}\\ &= \lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),\alpha\beta\big)\\ &= \lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),1\big)\cdot\psi(1,\alpha\beta).\end{aligned} \tag{3.9}$$

From the equalities (3.8) and (3.9), we get

$$i\big(f(\alpha,\beta),1\big)\cdot\lambda(\alpha\beta)^{-1} = \lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),1\big).$$

Conversely, suppose there exists $\lambda : Q \to C$ satisfying this equality. One simply defines $\theta : E \to E'$ by

$$\theta(a,\alpha) = i(a,1)\cdot\lambda(\alpha)^{-1}\cdot\psi(1,\alpha).$$

We claim that θ is a desired homomorphism. This is shown by a series of calculations:

$$
\begin{aligned}
&\theta(a,\alpha)\cdot\theta(b,\beta)\\
&\quad=\{i(a,1)\cdot\lambda(\alpha)^{-1}\cdot\psi(1,\alpha)\}\cdot\{i(b,1)\cdot\lambda(\beta)^{-1}\cdot\psi(1,\beta)\} \qquad (\text{definition of } \theta)\\
&\quad=i(a,1)\cdot\lambda(\alpha)^{-1}\cdot\{\psi(1,\alpha)\cdot i(b,1)\cdot\psi(1,\alpha)^{-1}\}\\
&\qquad\cdot\{\psi(1,\alpha)\cdot\lambda(\beta)^{-1}\cdot\psi(1,\alpha)^{-1}\}\cdot\{\psi(1,\alpha)\cdot\psi(1,\beta)\}\\
&\quad=i(a,1)\cdot\lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(i(b,1)\big)\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),\alpha\beta\big)\\
&\qquad\qquad\big(\text{since } \psi \text{ is a homomorphism and } (1,\alpha)(1,\beta)=\big(f(\alpha,\beta),\alpha\beta\big)\big)\\
&\quad=i(a,1)\cdot\tilde{\varphi}'(\alpha)\big(i(b,1)\big)\cdot\lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),1\big)\cdot\psi(1,\alpha\beta)\\
&\qquad\qquad\big(\lambda(\alpha)^{-1}\in C\big)\\
&\quad=i(a,1)\cdot i\big(\tilde{\varphi}(\alpha)(b,1)\big)\cdot\{\lambda(\alpha)^{-1}\cdot\tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta),1\big)\}\cdot\psi(1,\alpha\beta)\\
&\qquad\qquad\big(\tilde{\varphi}'(\alpha)\circ i=i\circ\tilde{\varphi}(\alpha)\big)\\
&\quad=i(a,1)\cdot i\big(\tilde{\varphi}(\alpha)(b,1)\big)\cdot\{i\big(f(\alpha,\beta),1\big)\cdot\lambda(\alpha\beta)^{-1}\}\cdot\psi(1,\alpha\beta)\\
&\qquad\qquad(\text{equality } (3.7))\\
&\quad=i\big(a\cdot\tilde{\varphi}(\alpha)(b)\cdot f(\alpha,\beta)\big)\cdot\lambda(\alpha\beta)^{-1}\cdot\psi(1,\alpha\beta)\\
&\quad=\theta\big(a\cdot\tilde{\varphi}(\alpha)(b)\cdot f(\alpha,\beta),\alpha\beta\big) \qquad (\text{definition of } \theta)\\
&\quad=\theta\big((a,\alpha)\cdot(b,\beta)\big) \qquad (\text{group operation in } E).
\end{aligned}
$$

(2) For $x\in E$, $c\in C=C_{G'}(i(G))$ and $a\in G$, it is easy to see that $\mu'(\theta_0(x))(c)$ and $i(a)$ commute with each other. This shows that the homomorphism $E\xrightarrow{\theta_0}E'\xrightarrow{\mu'}\mathrm{Aut}(G')$, where μ' is conjugation in E', leaves C invariant. Therefore we have a homomorphism

$$E\xrightarrow{\theta_0}E'\xrightarrow{\mu'}\mathrm{Aut}(C).$$

Furthermore, for $a\in G$,

$$(\mu'\circ\theta_0)(a)=\mu'\big(i(a)\big)$$

is trivial on C. Thus $E\to\mathrm{Aut}(C)$ factors through Q. We have obtained a homomorphism

$$\varphi''=\mu'\circ\theta_0: Q\to\mathrm{Aut}(C).$$

(Note that φ'' is actually a homomorphism, even though $\tilde{\varphi}: Q\to\mathrm{Aut}(G')$ is not.) Let $\theta: E\to E'$ be a homomorphism fitting the diagram. Then θ must be of the form

$$\theta(\alpha)=\eta(\alpha)\theta_0(\alpha)$$

for some map $\eta: E\to G'$. It is easy to verify that η satisfies

$$\eta(\alpha\beta)=\eta(\alpha)\cdot\theta_0(\alpha)\eta(\beta)\theta_0(\alpha)^{-1}=\eta(\alpha)\cdot\varphi''(\alpha)\big(\eta(\beta)\big).$$

Since $\theta|_G = i = \theta_0|_G$, we have $\eta(a) = 1$ for all $a \in G$. For any $a \in G$ and $\alpha \in E$, $\eta(\alpha a) = \eta(\alpha)$ from the above equality. Moreover, $\eta(\alpha) = \eta(\alpha \cdot (\alpha^{-1} a \alpha)) = \eta(a\alpha) = \eta(a) \cdot \theta_0(a)\eta(\alpha)\theta_0(a)^{-1} = 1 \cdot i(a)\eta(\alpha)i(a)^{-1} = i(a)\eta(\alpha)i(a)^{-1}$. This shows η has values in the centralizer $C = C_{G'}(i(G))$. Thus

$$\eta : Q \to C = C_{G'}\big(i(G)\big),$$

and hence $\eta \in Z^1_{\varphi''}(Q; C)$. See Section 3.6.

Conversely, a map $\eta : Q \to C$ satisfying the cocycle condition gives rise to a homomorphism $\eta \cdot \theta_0 : E \to E'$. Consequently,

$$\eta \longleftrightarrow \eta \cdot \theta_0$$

is a one-one correspondence between $Z^1_{\varphi''}(Q; C)$ and the set of all homomorphisms $E \to E'$ inducing i and $\bar{\theta}$ on G and Q, respectively.

Let $\eta, \eta' \in Z^1_{\varphi''}(Q; C)$, and suppose there exists $c \in C$ such that

$$(\eta' \cdot \theta_0)(\alpha) = c \cdot (\eta \cdot \theta_0)(\alpha) \cdot c^{-1}$$

for all $\alpha \in E$. This is equal to

$$\eta'(\alpha) = c \cdot \eta(\alpha) \cdot \varphi''(\alpha)\big(c^{-1}\big);$$

i.e.,

$$[\eta'] = [\eta] \in H^1_{\varphi''}(Q; C).$$

Consequently, $H^1_{\varphi''}(Q; C)$ classifies all homomorphisms $E \to E'$ inducing i and $\bar{\theta}$ on G and Q, respectively, up to conjugation by elements of C.

(3) Let

$$1 \to G' \to \mathcal{E} \to Q \to 1$$

be the pullback (see Section 3.3.2) of the exact sequence $1 \to G' \to E' \to Q' \to 1$ via $\bar{\theta} : Q \to Q'$. For any $1 \to G \to E'' \to Q \to 1$, there exists a homomorphism $E'' \to E'$ making the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E'' & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle i} & & \downarrow & & \downarrow{\scriptstyle \bar{\theta}} & & \\ 1 & \longrightarrow & G' & \longrightarrow & E' & \longrightarrow & Q' & \longrightarrow & 1 \end{array}$$

commutative if and only if there exists a homomorphism $E'' \to \mathcal{E}$ making the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E'' & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle i} & & \downarrow & & \downarrow{\scriptstyle \bar{\theta}} & & \\ 1 & \longrightarrow & G' & \longrightarrow & \mathcal{E} & \longrightarrow & Q' & \longrightarrow & 1 \end{array}$$

commutative. Suppose $\theta_0 : E \to E'$ exists as in the diagram (3.6). This yields $E \to \mathcal{E}$. We identify

$$\begin{aligned} \mathrm{Opext}(G, Q, \tilde{\varphi}) &\approx H^2(Q; \mathcal{Z}(G)), \\ \mathrm{Opext}(G, Q, \tilde{\varphi}') &\approx H^2(Q; \mathcal{Z}(G')). \end{aligned}$$

Since $i(\mathcal{Z}(G)) \subset \mathcal{Z}(G')$, $i : \mathcal{Z}(G) \to \mathcal{Z}(G')$ induces a homomorphism

$$i_* : H^2(Q; \mathcal{Z}(G)) \longrightarrow H^2(Q; \mathcal{Z}(G')).$$

Then, clearly, $[E''] \in \mathrm{Opext}(G, Q, \tilde{\varphi})$ can map into $[E']$ (or, equivalently into $\mathcal{E}$) if and only if $[E''] - [E] \in H^2(Q; \mathcal{Z}(G))$ maps to $0 \in H^2(Q; \mathcal{Z}(G'))$. □

Unless there is one definite extension E and a homomorphism θ_0, one cannot conclude that there is a homomorphism $E'' \to E'$ using this cohomology homomorphism. On the other hand, suppose that there is an extension E and a homomorphism $E \to E'$, and that $H^2(Q; \mathcal{Z}(G')) = 0$. Then every extension is in the kernel of $H^2(Q; \mathcal{Z}(G)) \to H^2(Q; \mathcal{Z}(G'))$, and hence every extension with the abstract kernel φ can be mapped into E'.

4. Seifert fiber space construction with $\mathrm{TOP}_G(G \times W)$

4.1. *Introduction*

Recall, from Lemma 2.5.3, that the group of all weakly G-equivariant self-homeomorphisms of $G \times W$ is

$$\mathrm{TOP}_G(G \times W) = \mathrm{M}(W, G) \rtimes (\mathrm{Aut}(G) \times \mathrm{TOP}(W)),$$

the universal group for a Seifert Construction.

Let $\Gamma \subset G$ be a discrete subgroup, $\rho : Q \to \mathrm{TOP}(W)$ a properly discontinuous action, and $1 \to \Gamma \to \pi \to Q \to 1$ a group extension (with abstract kernel $\varphi : Q \to \mathrm{Out}(G)$).

Our goal is to find a homomorphism $\theta : \pi \to \mathrm{TOP}_G(G \times W)$ which makes the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi \times \rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \rtimes \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

commutative.

The construction of a homomorphism $\theta : \pi \to \mathrm{TOP}_G(G \times W)$ will be achieved in two steps. First, we map π into a group $G\pi$ (see Section 3.3.1), and second $G\pi$ into $\mathrm{TOP}_G(G \times W)$.

For the first step, we require, in addition, that Γ be a lattice in a simply connected completely solvable Lie group or a semi-simple Lie group for which Mostow's rigidity theorem holds. Then we stack the two steps together to get our desired homomorphisms. A complete proof for the Abelian case is given. The other cases rely heavily on the Abelian case, and we sketch and reference the proofs for them.

As it turns out, the proofs of the existence also leads to uniqueness and rigidity theorems for these homomorphisms. The meaning of existence, uniqueness and rigidity are then explored, for it is these properties on which many of the applications of the Seifert Construction are based. Finally, we observe that the Main Theorem 4.3.2 of this section remains valid in the smooth category. We treat the second step first.

4.2. *Mapping an extension of G into* $\mathrm{TOP}_G(G \times W)$

Interpreting the algebraic criterion of Theorem 3.7.2 with $E = G\pi$ and $E' = \mathrm{TOP}_G(G \times W)$, we get

PROPOSITION 4.2.1. *Let $\rho : Q \to \mathrm{TOP}(W)$ be a properly discontinuous action, and $1 \to G \to E \to Q \to 1$ an extension with abstract kernel φ. Suppose $[f] \in \mathrm{Opext}(Q, G, \tilde{\varphi})$ represents the extension E. That is, $E = G \times_{(f,\tilde{\varphi})} Q$ (see Section 3.2.2). (Then $\tilde{\varphi} : Q \to \mathrm{Aut}(G)$ and $f : Q \times Q \to G$ satisfy the equalities (3.1), (3.2) and (3.3) of Section 3.2.)*

(1) *There exists a homomorphism $\theta : E \to \mathrm{TOP}_G(G \times W)$ making*

$$\begin{array}{ccccccccc}
1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi\times\rho} & & \\
1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1
\end{array}$$

commutative if and only if there exists a map $\lambda : Q \to \mathrm{M}(W, G)$ satisfying

$$f(\alpha, \beta) = \{\tilde{\varphi}(\alpha) \circ \lambda(\beta) \circ \rho(\alpha)^{-1}\} \cdot \lambda(\alpha) \cdot \lambda(\alpha\beta)^{-1} \in r(G) \subset \mathrm{M}(W, G) \qquad (4.1)$$

for all $\alpha, \beta \in Q$. [Compare this with the equality (3.4) defining $\delta\lambda$.]

(2) *Suppose there exists a homomorphism $\theta_0 : E \to \mathrm{TOP}_G(G \times W)$ as above. Then $\mu' \circ \theta_0$* (μ' is conjugation in $\mathrm{TOP}_G(G \times W)$) *induces a homomorphism $\varphi'' : Q \to \mathrm{Aut}(\mathrm{M}(W, G))$.*

The set of all θ's fitting the diagram, up to conjugation by elements of $\mathrm{M}(W, G)$, is in one-one correspondence with $H^1_{\varphi''}(Q; \mathrm{M}(W, G))$.

(3) *Suppose that there exists a homomorphism $\theta_0 : E \to \mathrm{TOP}_G(G \times W)$ as above. Then $\theta_0(\mathcal{Z}(G)) \subset \mathrm{M}(W, \mathcal{Z}(G))$. Let $[E''] \in \mathrm{Opext}(Q, G, \tilde{\varphi})$ be an extension. Then there exist a homomorphism $E'' \to \mathrm{TOP}_G(G \times W)$ completing the diagram if and only if $[E''] - [E]$ lies in the kernel of $\ell_* : H^2(Q; \mathcal{Z}(G)) \to H^2(Q; \mathrm{M}(W, \mathcal{Z}(G)))$.*

PROOF. (1) We apply Theorem 3.7.2(1) to the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & & & \downarrow{\scriptstyle \varphi\times\rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W,G)\rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G\times W) & \longrightarrow & \mathrm{Out}(G)\times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

Let $\psi : E \to \mathrm{TOP}_G(G \times W)$ be given by

$$\psi(\tilde{\alpha}) = \big(1, \tilde{\varphi}(\tilde{\alpha}), \rho(\tilde{\alpha})\big) \in \mathrm{M}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{TOP}(W)\big),$$

where $\tilde{\varphi}(\tilde{\alpha}) \in \mathrm{Aut}(G)$ is conjugation in E, and $\rho : E \to Q \to \mathrm{TOP}(W)$ (abuse of notation) is the given properly discontinuous action of Q on W. In particular,

$$\psi(1, \alpha) = \big(1, \tilde{\varphi}(\alpha), \rho(\alpha)\big).$$

Conditions (A) in Section 3.7 is easily checked. For (B), Let $a \in G$ and $\tilde{\alpha} \in E$. Then

$$\begin{aligned} \ell\big(\tilde{\alpha} \cdot a \cdot \tilde{\alpha}^{-1}\big) &= \ell\big(\tilde{\varphi}(\tilde{\alpha})(a)\big) \\ &= \big(\tilde{\varphi}(\tilde{\alpha})(a)^{-1}, \mu\big(\tilde{\varphi}(\tilde{\alpha})(a)\big), 1\big). \end{aligned}$$

On the other hand,

$$\begin{aligned} \psi(\tilde{\alpha}) \cdot \ell(a) \cdot \psi(\tilde{\alpha})^{-1} &= \big(1, \tilde{\varphi}(\tilde{\alpha}), \rho(\tilde{\alpha})\big) \cdot \big(a^{-1}, \mu(a), 1\big) \cdot \big(1, \tilde{\varphi}(\tilde{\alpha})^{-1}, \rho(\tilde{\alpha})^{-1}\big) \\ &= \big(\tilde{\varphi}(\tilde{\alpha})(a)^{-1}, \mu\big(\tilde{\varphi}(\tilde{\alpha})(a)\big), 1\big). \end{aligned}$$

Thus,

$$\ell\big(\tilde{\alpha} \cdot a \cdot \tilde{\alpha}^{-1}\big) = \psi(\tilde{\alpha}) \cdot \ell(a) \cdot \psi(\tilde{\alpha})^{-1}$$

as we wanted.

Now we examine the equality (3.7).

$$\begin{aligned} \ell\big(f(\alpha, \beta)\big) \cdot \lambda(\alpha\beta)^{-1} &= \big(f(\alpha, \beta)^{-1}, \mu\big(f(\alpha, \beta)\big), 1\big) \cdot \big(\lambda(\alpha\beta)^{-1}, 1, 1\big) \\ &= \big(\lambda(\alpha\beta)^{-1} \cdot f(\alpha, \beta)^{-1}, \mu\big(f(\alpha, \beta)\big), 1\big), \end{aligned}$$

while

$$\begin{aligned} &\lambda(\alpha)^{-1} \cdot \tilde{\varphi}'(\alpha)\big(\lambda(\beta)^{-1}\big) \cdot \psi\big(f(\alpha, \beta)\big) \\ &\quad = \big(\lambda(\alpha)^{-1}, 1, 1\big) \cdot \big\{\psi(1, \alpha) \cdot \lambda(\beta)^{-1} \cdot \psi(1, \alpha)^{-1}\big\} \cdot \big(1, \mu\big(f(\alpha, \beta)\big), 1\big) \\ &\quad = \big(\lambda(\alpha)^{-1}, 1, 1\big) \cdot \big\{\big(1, \tilde{\varphi}(\alpha), \rho(\alpha)\big) \cdot \big(\lambda(\beta)^{-1}, 1, 1\big) \cdot \big(1, \tilde{\varphi}(\alpha), \rho(\alpha)\big)^{-1}\big\} \\ &\qquad \cdot \big(1, \mu\big(f(\alpha, \beta)\big), 1\big) \\ &\quad = \big(\lambda(\alpha)^{-1}, 1, 1\big) \cdot \big(\tilde{\varphi}(\alpha) \circ \lambda(\beta)^{-1} \circ \rho(\alpha)^{-1}, 1, 1\big) \cdot \big(1, \mu\big(f(\alpha, \beta)\big), 1\big) \\ &\quad = \big(\lambda(\alpha)^{-1} \cdot \big\{\tilde{\varphi}(\alpha) \circ \lambda(\beta)^{-1} \circ \rho(\alpha)^{-1}\big\}, \mu\big(f(\alpha, \beta)\big), 1\big). \end{aligned}$$

Therefore,

$$\ell(f(\alpha,\beta))\cdot\lambda(\alpha\beta)^{-1}=\lambda(\alpha)^{-1}\cdot\tilde{\varphi}(\alpha)\big(\lambda(\beta)^{-1}\big)\cdot\psi\big(f(\alpha,\beta)\big)$$

if and only if

$$\lambda(\alpha\beta)^{-1}\cdot f(\alpha,\beta)^{-1}=\lambda(\alpha)^{-1}\cdot\big\{\tilde{\varphi}(\alpha)\circ\lambda(\beta)^{-1}\circ\rho(\alpha)^{-1}\big\}$$

or equivalently,

$$f(\alpha,\beta)=\big\{\tilde{\varphi}(\alpha)\circ\lambda(\beta)\circ\rho(\alpha)^{-1}\big\}\cdot\lambda(\alpha)\cdot\lambda(\alpha\beta)^{-1}$$

as elements of G. This is the identity (4.1).

(2) Suppose there exists a homomorphism $\theta_0 : E \to \mathrm{TOP}_G(G\times W)$ as above. Since $\ell(G)$ is normal in $\mathrm{TOP}_G(G\times W)$, and the centralizer of $\ell(G)$ in $\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)$ is $\mathrm{M}(W,G)$, $\mathrm{M}(W,G)$ is normal in $\mathrm{TOP}_G(G\times W)$. Consequently, conjugation by elements of $\mathrm{TOP}_G(G\times W)$ leaves $\mathrm{M}(W,G)$ invariant, and $\mu'\circ\theta_0$ induces a homomorphism $\varphi'' : Q\to \mathrm{Aut}(\mathrm{M}(W,G))$. The rest is clear.

(3) Observe that the center of $\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)$ is $\mathrm{M}(W,\mathcal{Z}(G))$ so that $\ell(\mathcal{Z}(G))\subset\mathrm{M}(W,\mathcal{Z}(G))$. Now apply Theorem 3.7.2(3). □

REMARK 4.2.2. When the condition (4.1) is satisfied, vanishing of $H^1(Q;\mathrm{M}(W,G))$ and $H^2(Q;\mathrm{M}(W,\mathcal{Z}(G)))$ (e.g., if $\mathcal{Z}(G)$ is contractible) yields strong consequences by the statements (2) and (3). Namely,

(2) The homomorphism θ fitting the diagram is unique up to conjugation by elements of $\mathrm{M}(W,G)$.

(3) Every extension with the abstract kernel φ can be mapped into $\mathrm{TOP}_G(G\times W)$.

The vanishing of $H^2(Q;\mathrm{M}(W,\mathcal{Z}(G)))$ only says that there is at most one extension of $\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)$ by Q for a given abstract kernel. This alone is not necessarily enough to guarantee that the condition (4.1) is automatically satisfied because the coefficient groups G and $\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)$ have distinct "operator groups". On the other hand, we have the following important observation.

THEOREM 4.2.3. *If $\tilde{\varphi}_C : Q \xrightarrow{\tilde{\varphi}} \mathrm{Aut}(\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G))\to\mathrm{Aut}(\mathrm{M}(W,G))$ is a homomorphism, and $H^2(Q;\mathrm{M}(W,\mathcal{Z}(G)))=0$, then θ exists as in Proposition 4.2.1, and consequently condition (4.1) is satisfied.*

PROOF. Let

$$1\longrightarrow \mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)\longrightarrow\mathcal{E}\longrightarrow Q\longrightarrow 1$$

be the pullback of $1\longrightarrow\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)\longrightarrow\mathrm{M}(W,G)\rtimes(\mathrm{Aut}(G)\times\mathrm{TOP}(W))\longrightarrow \mathrm{Out}(G)\times\mathrm{TOP}(W)\longrightarrow 1$. Then

$$\mathcal{E}=\big(\mathrm{M}(W,G)\rtimes\mathrm{Inn}(G)\big)\times_{(h,\tilde{\varphi})}Q$$

for some maps $\tilde{\varphi}: Q \to \mathrm{Aut}(\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G))$ and $h: Q \times Q \to \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ satisfying

$$\tilde{\varphi}(\alpha) \circ \tilde{\varphi}(\beta) = \mu\big(h(\alpha, \beta)\big) \circ \tilde{\varphi}(\alpha\beta),$$
$$h(\alpha, 1) = 1 = h(1, \beta),$$
$$h(\alpha, \beta) \cdot h(\alpha\beta, \gamma) = \tilde{\varphi}(\alpha)\big(h(\beta, \gamma)\big) \cdot h(\alpha, \beta\gamma)$$

for every $\alpha, \beta, \gamma \in Q$. Note that the pair $(h, \tilde{\varphi})$ is not unique. (In fact, each $\tilde{\varphi}(\alpha)$ can change by conjugation by an element of $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$, and h changes accordingly.) The map $\tilde{\varphi}$ induces two maps $\tilde{\varphi}_G : Q \to \mathrm{Aut}(G)$ and $\tilde{\varphi}_C : Q \to \mathrm{Aut}(\mathrm{M}(W, G))$. The second map in the definition of $\tilde{\varphi}_C$ is obtained as follows. Since $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) = \ell(G) \times_{\mathcal{Z}(G)} \mathrm{M}(W, G)$, and $\ell(G)$ is normal in $\mathrm{TOP}_G(G \times W)$, an automorphism of $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ leaving $\ell(G)$ invariant induces an automorphism of $\mathrm{M}(W, G)$. Also $\tilde{\varphi}_G$ induces a homomorphism $\varphi: Q \to \mathrm{Out}(G)$.

Now let

$$1 \to G \to E \to Q \to 1$$

be any extension with abstract kernel $\varphi: Q \to \mathrm{Out}(G)$. Then there exists a map $f: Q \times Q \to G$ such that $E = G \times_{(f, \tilde{\varphi}_G)} Q$ satisfies

$$\tilde{\varphi}_G(\alpha) \circ \tilde{\varphi}_G(\beta) = \mu\big(f(\alpha, \beta)\big) \circ \tilde{\varphi}_G(\alpha\beta),$$
$$f(\alpha, 1) = 1 = f(1, \beta),$$
$$f(\alpha, \beta) \cdot f(\alpha\beta, \gamma) = \tilde{\varphi}_G(\alpha)\big(f(\beta, \gamma)\big) \cdot f(\alpha, \beta\gamma)$$

for every $\alpha, \beta, \gamma \in Q$. Let us examine the first equality. This equality holds in the group $\mathrm{Aut}(G)$. Now $f(\alpha, \beta) \in \ell(G)$ and $\mathrm{M}(W, G) = C$ centralizes $\ell(G)$. Hence $\mu(f(\alpha, \beta))$ on C is trivial. If we replace $\tilde{\varphi}_G$ in the first equation by $\tilde{\varphi}_C$, then the equation is satisfied on $\mathrm{M}(W, G)$ in $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ because $\mu(f(\alpha, \beta))$ is trivial and $\tilde{\varphi}_C$ is a homomorphism. Since $\ell(G)$ and $\mathrm{M}(W, G)$ generate $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$, the first equality still holds in $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ if we replace $\tilde{\varphi}_G$ with $\tilde{\varphi}$. Because $\tilde{\varphi}|_{\ell(G)}(\alpha)(\ell(g)) = \ell(\tilde{\varphi}_G(\alpha)(g))$, the remaining equalities hold inside $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ if we interpret $f: Q \times Q \to G$ as $f: Q \times Q \to \ell(G) \subset \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$. Consequently, $E = G \times_{(f, \tilde{\varphi}_G)} Q$ is sitting in $\mathcal{E}' = (\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)) \times_{(f, \tilde{\varphi})} Q$.

Now, if $H^2(Q; M(W, \mathcal{Z}(G))) = 0$, the extensions $\mathcal{E}$ and $\mathcal{E}'$ are congruent, and we have a desired homomorphism $E \to \mathcal{E}$. This theorem will be used to prove Theorem 4.3.2. □

4.3. *Construction with* $\mathrm{TOP}_G(G \times W)$

4.3.1. We shall say a discrete group Γ is *special* if Γ is isomorphic to a lattice in any one of the following Lie groups G:

(S1) $\mathbb{R}^k$ for some $k > 0$,

(S2) a simply connected nilpotent Lie group, (S3) a simply connected completely solvable Lie group; that is, for each $X \in \mathcal{G}$, the Lie algebra of G, $\mathrm{ad}(X) : \mathcal{G} \to \mathcal{G}$ has only real eigenvalues,

(S4) a semi-simple centerless Lie group without any normal compact factors and if G contains any 3-dimensional factors (i.e., $\mathrm{PSL}(2, \mathbb{R})$), then the projection of the lattice to each of these factors is dense.

We shall also call the Lie group G *special*. Such groups possess ULIEP (Unique Lattice Isomorphism Extension Property). That is, any isomorphism between such lattices extends uniquely to an isomorphism of G. Even more generally, if Γ_i is a lattice in a special G_i $(i = 1, 2)$, then any isomorphism $\varphi : \Gamma_1 \to \Gamma_2$ extends uniquely to an isomorphism $\Psi : G_1 \to G_2$.

Notice that: type (S1) $\Rightarrow$ type (S2) $\Rightarrow$ type (S3).

The following is the main construction. Its proof is deferred until Section 4.5.

THEOREM 4.3.2 ([12,26,31,54]). *Let $\Gamma \subset G$ be a* special *lattice. Let $\rho : Q \to \mathrm{TOP}(W)$ be a properly discontinuous action. Then for any extension $1 \to \Gamma \to \pi \to Q \to 1$ (with abstract kernel $\varphi : Q \to \mathrm{Out}(G)$) the following are true*:

(1) Existence: *There exists a proper action $\theta : \pi \to \mathrm{TOP}_G(G \times W)$ making the diagram*

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi \times \rho} & & \\
1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \rtimes \mathrm{TOP}(W) & \longrightarrow & 1
\end{array} \tag{4.2}$$

commutative. For completely solvable G we have in addition:

(2) Uniqueness: *Congruent extensions are conjugate in* $\mathrm{M}(W, G) \subset \mathrm{TOP}_G(G \times W)$. *More precisely, suppose $\theta_1, \theta_2 : \pi \to \mathrm{TOP}_G(G \times W)$ are two homomorphisms which fit in diagram* (4.2) *with fixed ℓ and $\varphi \times \rho$, then there exists $\lambda \in \mathrm{M}(W, G)$ such that $\theta_2 = \mu(\lambda) \circ \theta_1$.*

(3) Rigidity: *Suppose $\theta_1, \theta_2 : \pi \to \mathrm{TOP}_G(G \times W)$ are two homomorphisms which fit in the diagram* (4.2) (possibly with distinct ℓ and ρ). *Then there exists $(\lambda, a, h) \in \mathrm{TOP}_G(G \times W)$ such that $\theta_2 = \mu(\lambda, a, h) \circ \theta_1$, provided that $\rho_2 = \mu(h) \circ \rho_1$.*

Uniqueness and rigidity for the semi-simple case (S4) will be discussed in Sections 4.5.6 and 4.5.7

4.4. *The meaning of existence, uniqueness and rigidity*

4.4.1. (Existence) Suppose there exists $\theta : \pi \to \mathrm{TOP}_G(G \times W)$ making the diagram (4.2) commutative. Then $\theta(\pi)$ acts properly on $G \times W$. This yields the injective Seifert fibering

$$\ell(\Gamma) \setminus \ell(G) \longrightarrow \theta(\pi) \setminus (G \times W) \xrightarrow{p} Q \setminus W$$

with the typical fiber $\Gamma \setminus G$ and base $Q \setminus W$. We have explained the properties of such a fibering, and its singular fibers in Sections 2.7, 2.8 and 2.9.

4.4.2. (Uniqueness) Let θ_0, θ_1 be homomorphisms of π into $\mathrm{TOP}_G(G \times W)$ such that

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \big\downarrow{\scriptstyle \ell} & & \big\downarrow{\scriptstyle \theta_i} & & \big\downarrow{\scriptstyle \phi \times \rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

is commutative. If there exists $\lambda \in \mathrm{M}(W, G)$ such that $\theta_1 = \mu(\lambda) \circ \theta_0$, the map $\lambda : W \to G$ induces a homeomorphism of $G \times W$ by

$$\lambda(x, w) = \bigl(x \cdot (\lambda(w))^{-1}, w\bigr),$$

see Section 2.5. This, in turn, yields a homeomorphism $[\lambda] : \frac{G \times W}{\theta_0(\pi)} \to \frac{G \times W}{\theta_1(\pi)}$. As the commuting diagram shows

$$\begin{array}{ccc} M_0 = \frac{G \times W}{\theta_0(\pi)} & \xrightarrow{[\lambda]} & \frac{G \times W}{\theta_1(\pi)} = M_1 \\ \big\downarrow & & \big\downarrow \\ \rho(Q) \backslash W & \xrightarrow{=} & \rho(Q) \backslash W \end{array}$$

the map $[\lambda]$ induces the identity map on the base space. In fact, λ is G-equivariant (that is, it commutes with the left principal G-action). Such spaces M_0 and M_1 are said to be *strictly equivalent.*

When G is special of type (S1), (S2) or (S3); or W is contractible, of type (S4) (see Section 4.3.1), λ is homotopic to the constant map $e : W \to G$ (e is the identity element of G). Then the path $\{\lambda_t : 0 \leqslant t \leqslant 1\}$ gives rise to a continuous family of homomorphisms $\theta_t : \pi \to \mathrm{TOP}_G(G \times W)$ by

$$\theta_t(\tilde{\alpha}) = \lambda_t \cdot \theta_0(\tilde{\alpha}) \cdot \lambda_t^{-1}$$

for $\tilde{\alpha} \in \pi$, and consequently a continuous family of homeomorphisms

$$[\theta_t] : \frac{G \times W}{\theta_0(\pi)} \longrightarrow \frac{G \times W}{\theta_t(\pi)}.$$

Thus $\frac{G \times W}{\theta_0(\pi)}$ can be deformed to $\frac{G \times W}{\theta_1(\pi)}$ by moving just along the fibers. In fact, the family $\lambda_t : G \times W \to G \times W$ is G-equivariant. If $\theta_0(\pi)$ commutes with $\ell(G)$, then the deformation $[\theta_t]$ is G-equivariant.

4.4.3. (Rigidity) Let θ_0, θ_1 be homomorphisms of π into $\mathrm{TOP}_G(G \times W)$ such that

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \big\downarrow{\scriptstyle \ell_i} & & \big\downarrow{\scriptstyle \theta_i} & & \big\downarrow{\scriptstyle \phi_i \times \rho_i} & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

is commutative. If there exists $(\lambda, a, h) \in \mathrm{TOP}_G(G \times W)$ such that $\theta_1 = \mu(\lambda, a, h) \circ \theta_0$, the map $(\lambda, a, h): G \times W \to G \times W$ induces a homeomorphism $[\lambda, a, h]$

$$\begin{array}{ccc} M_0 = \frac{G \times W}{\theta_0(\pi)} & \xrightarrow{[\lambda,a,h]} & \frac{G \times W}{\theta_1(\pi)} = M_1 \\ \downarrow & & \downarrow \\ \rho_0(Q) \backslash W & \xrightarrow{\bar{h}} & \rho_1(Q) \backslash W \end{array}$$

which preserves the fibers. More precisely, $[\lambda, a, h]$ is defined by

$$[\lambda, a, h]\big([x, w]\big) = \big[(\lambda, a, h)(x, w)\big].$$

Since $(\lambda, a, h) \circ \theta_0(\alpha) = \theta_1(\alpha) \circ (\lambda, a, h)$ for all $\alpha \in \pi$, the map $[\lambda, a, h]$ is well-defined. Further, $(\lambda, a, h)(x, w) = (a(x) \cdot \lambda h(w), h(w))$ shows that $[\lambda, a, h]$ is a map from M_0 to M_1. That is, $[\lambda, a, h]$ is the descent of the weakly equivariant fiber preserving map (λ, a, h) sending G fibers to G fibers. Such spaces M_0 and M_1 are said to be *equivalent*.

The conjugation of θ_0 by (λ, a, h) is called a *Seifert automorphism* of $G \times W$. The induced homeomorphism $[\lambda, a, h]: M_0 \to M_1$ is called a *Seifert equivalence* or *Seifert automorphism*.

If Γ is characteristic in π, any automorphism of π induces an automorphism of Γ and Q. Consequently, if M_0 and M_1 are Seifert fiber spaces modelled on $G \times W$ which have the same fundamental group (or "orbifold fundamental group") they are equivalent as Seifert fiber spaces, provided that the base spaces are rigidly related (i.e., there exists $h \in \mathrm{TOP}(W)$ for which $\rho_1 = \mu(h) \circ \rho_0$).

Similarly, if $\tau: \pi \to \pi'$ is an isomorphism carrying Γ isomorphically onto Γ', and inducing an isomorphism $\bar{\tau}: Q \to Q'$, then the analogous rigidity statement for $M = \theta(\pi) \backslash (G \times W)$ and $M' = \theta'(\pi') \backslash (G \times W)$ holds, cf. [12, 8.5] and [26, 2.4]. More precisely, if there exists $h \in \mathrm{TOP}(W)$ such that $\mu(h) \circ \rho = \rho' \circ \bar{\tau}$, then there exists $\hat{h} = (\lambda, a, h) \in \mathrm{TOP}_G(G \times W)$ such that $\theta' \circ \tau = \mu(\hat{h}) \circ \theta$.

EXAMPLE 4.4.4 (*When W is a point*). Suppose $W = \{p\}$, a point. Then

$$\begin{aligned} \mathrm{TOP}_G(G \times W) &= \mathrm{M}(p, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{TOP}(p)\big) \\ &= r(G) \rtimes \mathrm{Aut}(G) \\ &= \ell(G) \rtimes \mathrm{Aut}(G). \end{aligned}$$

Also, note that $\mathrm{M}(p, G) \rtimes \mathrm{Inn}(G) = r(G) \rtimes \mathrm{Inn}(G) = \ell(G) \rtimes \mathrm{Inn}(G)$. Since Q acts on W properly, Q must be a finite group.

We assume that $\Gamma \subset G$ is a *special* lattice, and apply Theorem 4.3.2. Then diagram (4.2) becomes

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow \ell & & \downarrow \theta & & \downarrow \varphi & & \\ 1 & \longrightarrow & \ell(G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \ell(G) \rtimes \mathrm{Aut}(G) & \longrightarrow & \mathrm{Out}(G) & \longrightarrow & 1 \end{array}$$

Furthermore, since $\ell(\Gamma) \subset \ell(G)$, the above diagram induces

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \tilde{\varphi}} & & \\ 1 & \longrightarrow & \ell(G) & \longrightarrow & \ell(G) \rtimes \mathrm{Aut}(G) & \longrightarrow & \mathrm{Aut}(G) & \longrightarrow & 1 \end{array}$$

where $\tilde{\varphi}$ is a homomorphism lifting φ. [In particular, this shows that, for special G, the abstract kernel $\varphi: Q \to \mathrm{Out}(G)$ (with Q finite) lifts to a homomorphism $\tilde{\varphi}: Q \to \mathrm{Aut}(G)$.]

Consequently, $\theta(\pi) \subset \ell(G) \rtimes C$ for a finite subgroup $C \subset \mathrm{Aut}(G)$. This means that there exists a Riemannian metric on G for which $\theta(\pi) \subset \mathrm{Isom}(G)$.

Assume that θ is injective. (Recall that the kernel of θ is finite. Also, note that if π is torsion free, θ is injective.) For $\mathbb{R}^n = G$, $\tilde{\varphi}(Q)$ lies in a compact subgroup of $\mathrm{GL}(n, \mathbb{R})$ and can be conjugated within $\mathrm{GL}(n, \mathbb{R})$ to $\mathrm{O}(n)$ and consequently, by rigidity, $\theta(\pi)$ is conjugated by an affine diffeomorphism into the group of Euclidean motions $E(n) = \mathbb{R}^n \rtimes \mathrm{O}(n)$. An injective Seifert fiber space modelled on $\mathbb{R}^n \times$ point is a flat manifold if π is torsion free, and a flat orbifold otherwise. Similarly, if G is a connected, simply connected nilpotent Lie group, an injective Seifert fiber space modelled on $G \times$ point is an infra-nilmanifold if π is torsion free, and an infranil-orbifold otherwise. (Moreover, in both cases, each such manifold (or orbifold) must arise in this fashion.) Since W is a point, the rigidity on W always holds, and by the rigidity theorem, any two infra-G-manifolds modelled on special G with isomorphic fundamental groups are "affinely" diffeomorphic. (Recall that $\mathrm{TOP}_G(G \times W)$ reduces to $G \rtimes \mathrm{Aut}(G)$ in this case.) This generalizes the classical result of Bieberbach's from crystallographic groups to almost crystallographic groups in the nilpotent case.

4.5. *Proofs of Theorem 4.3.2*

There are two steps. First we map π into an extension E of G by Q, and second we map E into $\mathrm{TOP}_G(G \times W)$.

4.5.1. Every special Lie group G has the ULIEP. For every G with ULIEP, let $1 \to \Gamma \to \pi \to Q \to 1$ be an extension which represents $f \in \mathrm{Opext}(\Gamma, Q, \tilde{\varphi})$ satisfying (3.1), (3.2) and (3.3) of Section 3.2. One can form a pushout by the inclusion $i: \Gamma \hookrightarrow G$ to obtain $E = G\pi$. See Section 3.3.1. Then $f \in \mathrm{Opext}(G, Q, \tilde{\varphi})$ and the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 1 & \longrightarrow & G & \longrightarrow & E = G\pi & \longrightarrow & Q & \longrightarrow & 1 \end{array}$$

is commutative.

4.5.2. Now in order to map E into $\mathrm{TOP}_G(G \times W)$, we apply Proposition 4.2.1 for the cases (S1), (S2), and (S3). These cases depend upon the Abelian case whose proof we give in complete detail. When $G = \mathbb{R}^k$ is Abelian, the existence of λ in the equality (4.1)

$$f(\alpha, \beta) = \big(\tilde{\varphi}(\alpha) \circ \lambda(\beta) \circ \rho(\alpha)^{-1}\big) \cdot \lambda(\alpha) \cdot \lambda(\alpha\beta)^{-1}$$

means that $[f]$ will have to be 0 in $H^2(Q; \mathrm{M}(W, \mathbb{R}^k))$. The following lemma takes care of this problem:

LEMMA 4.5.3 ([12, Lemma 8.4] and [13, Theorem 7.1]). *Let $\rho: Q \to W$ be a properly discontinuous action with $Q \setminus W$ paracompact. For any homomorphism $\varphi: Q \to \mathrm{GL}(k, \mathbb{R})$, and an action of Q on $\mathrm{M}(W, \mathbb{R}^k)$ by $\alpha \cdot \lambda = \varphi(\alpha) \circ \lambda \circ \rho(\alpha)^{-1}$, $H^i(Q; \mathrm{M}(W, \mathbb{R}^k)) = 0$ for all $i > 0$.*

PROOF. We shall prove the lemma, as was done in [12], for $Q \setminus W$ compact. For the general case we refer the reader to [13]. Let U_x be a neighborhood of $x \in W$ with the property that $Q_x U_x = U_x$ and such that if $U_x \cap \alpha U_x \neq \emptyset$, then $\alpha \in Q_x$. Let $V_x = QU_x$. Take $M(V_x) \subset \mathrm{M}(W, \mathbb{R}^k)$ to be the submodule of all maps with support in V_x. Similarly, $M(U_x) \subset M(V_x) \subset \mathrm{M}(W, \mathbb{R}^k)$ is the subspace of all maps with support in U_x, and $M(U_x)$ is a Q_x-module. Then

$$\mathrm{Hom}_{\mathbb{Z}Q_x}\big(\mathbb{Z}Q; M(U_x)\big) \approx M(V_x)$$

because $\mathbb{Z}Q$ is a free $\mathbb{Z}Q_x$ module with a basis given by choosing coset representatives. Using Shapiro's lemma, it follows that

$$H^*\big(Q_x; M(U_x)\big) \approx H^*\big(Q; M(V_x)\big).$$

(See, e.g., [11, Chapter X, Proposition 7.4] or [9, Chapter III 5.8 and 6.2].) Now $H^i(Q_x; M(U_x)) = 0$ for all $i \geqslant 1$, because Q_x is a finite group and $M(U_x)$ is an $\mathbb{R}$-vector space.

If $y \notin Q(x)$, then neighborhoods U_x and U'_y can be chosen so that $V_x \cap U'_y = \emptyset$ because Q acts properly discontinuously on W. This insures that $Q \setminus W$ is Hausdorff (and completely regular). Because $Q \setminus W$ is compact, there exists a finite set of points $x_1, x_2, \ldots, x_n$ with neighborhoods $U_{x_1}, U_{x_2}, \ldots, U_{x_n}$ as above whose images, $V^*_{x_1}, V^*_{x_2}, \ldots, V^*_{x_n}$ in $Q \setminus W$ cover $Q \setminus W$. Take a partition of unity subordinate to this open covering. By composing each partition function with the quotient map $W \xrightarrow{p} Q \setminus W$, we get a partition of unity $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ on W subordinate to $V_1, V_2, \ldots, V_n$, where $V_i = p^{-1}(V^*_{x_i}) = Q_{x_i}(U_{x_i})$. Each partition function satisfies $\varepsilon_j(\alpha w) = \varepsilon_j(w)$, for all $\alpha \in Q$, $w \in W$. Now multiplication by ε_j is a Q-module homomorphism of $\mathrm{M}(W, \mathbb{R}^k)$

into $M(V_j)$. Consequently, there is induced homomorphisms ${\varepsilon_j}_*$ and the commutative diagram

$$\begin{array}{ccc} H^*(Q;\mathrm{M}(W,\mathbb{R}^k)) & \xrightarrow{\varepsilon_{j*}} & H^*(Q;\mathrm{M}(W,\mathbb{R}^k)) \\ \downarrow{\scriptstyle \varepsilon_{j*}} & & \uparrow{\scriptstyle i_*} \\ H^*(Q;M(V_j)) & \xrightarrow{=} & H^*(Q;M(V_j)) \end{array} \tag{4.3}$$

with $\varepsilon_{1*}+\varepsilon_{2*}+\cdots+\varepsilon_{n*}=\text{identity}$. This yields the lemma. □

4.5.4. For a nilpotent Lie group G, one shows that the equation

$$f(\alpha,\beta)=\left(\tilde{\varphi}(\alpha)\circ\lambda(\beta)\circ\rho(\alpha)^{-1}\right)\cdot\lambda(\alpha)\cdot\lambda(\alpha\beta)^{-1}$$

has a solution for λ. One reduces this problem to iterated applications of $H^2(Q;\mathrm{M}(W,\mathbb{R}^k)) = 0$ to guarantee the existence while $H^1(Q;\mathrm{M}(W,\mathbb{R}^k))=0$ for the uniqueness. The readers are referred to [26] for the details in the nilpotent case.

4.5.5. Theorem 4.2.3 is used to prove the existence in the completely solvable and type (S4) cases as well as providing a different procedure to that used in [26] for the nilpotent case. Details can be found in [41, Theorem 3]. The uniqueness for the completely solvable case is not given in [41] but it can be obtained inductively from the nilpotent case.

Rigidity in the abelian case was first proved by a cohomological argument, [12, §8.5]. In the nilpotent case, rigidity follows from uniqueness and ULIEP, [26, §2.4]. In fact, the same argument works for completely solvable G.

4.5.6. Suppose $\mathrm{Aut}(G)$ splits as $\mathrm{Inn}(G)\rtimes\mathrm{Out}(G)$, and $\mathcal{Z}(G)=0$ (for example, G of type (S4)). If we replace $\mathrm{M}(W,G)$ by $r(G)$ and $\mathrm{TOP}_G(G\times W)$ by $r(G)\rtimes(\mathrm{Aut}(G)\times\mathrm{TOP}(W))\subset\mathrm{TOP}_G(G\times W)$, then a homomorphism $\theta: E\to r(G)\rtimes(\mathrm{Aut}(G)\times\mathrm{TOP}(W))$ exists (with $\theta|_G=\ell$ on G) by Theorem 4.2.3. Thus E maps into

$$r(G)\rtimes\left(\mathrm{Aut}(G)\times\mathrm{TOP}(W)\right)=\left(\ell(G)\rtimes\mathrm{Aut}(G)\right)\times\mathrm{TOP}(W).$$

This induces two homomorphisms $E\to\ell(G)\rtimes\mathrm{Aut}(G)$ and $E\to\mathrm{TOP}(W)$. Consequently, the action of E on $G\times W$ is diagonal.

Note that uniqueness does not hold in $(\ell(G)\rtimes\mathrm{Aut}(G))\times\mathrm{TOP}(W)$ for $\Gamma\times Q$. The different strict equivalence classes are in 1-1 correspondence with $H^1(Q;r(G))=$ conjugacy classes of representations of Q into $r(G)$.

4.5.7. In [54], yet another construction is given in the (S4) case. Instead of $G\times W$ being the uniformizing space, $G/K\times W$ is chosen, where K is a maximal compact subgroup of G. Consequently, G/K is a simply connected symmetric space. The typical fiber becomes the locally symmetric space $F=\Gamma\setminus G/K$ and singular fiber is the quotient of F by a *finite group of isometries.* The space E has a finite covering E' (perhaps branched) where E' is $F \times (Q' \setminus W)$ and Q' is the kernel of $Q \to \mathrm{Out}(\Gamma)$. Since $\mathrm{Out}(\Gamma)$ is finite, Q' has finite index in Q and the finite group Q/Q' acts diagonally on $F \times (Q' \setminus W)$.

In [42], this is explained in more detail by showing that the uniformizing group $\mathrm{TOP}_{(G,K)}(G/K \times W)$ is isomorphic to $\mathrm{Isom}(G/K) \times \mathrm{TOP}(W)$. In this context, existence, uniqueness and rigidity are all shown to be valid. Typical fibers which are locally symmetric spaces have a great deal of geometric interest and the uniqueness and rigidity of this modified construction has significant applications. When modeling on $G/K \times W$ instead of $G \times W$ we shall call this the modified injective Seifert Construction of type (S4).

4.5.8. Now we stack the two constructions together to obtain a homomorphism $\Gamma \to \mathrm{TOP}_G(G \times W)$:

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & G & \longrightarrow & E & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow & & \\
1 & \longrightarrow & \mathrm{M}(W,G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1
\end{array}$$

This completes the proof of Theorem 4.3.2. □

4.6. *Smooth case*

If W is a smooth manifold and Q acts on it smoothly so that $\rho: Q \to \mathrm{Diff}(W)$, then the construction can be done smoothly. The universal group is now $\mathcal{C}(W, G) \rtimes (\mathrm{Aut}(G) \times \mathrm{Diff}(W))$ which is certainly the subgroup of weakly G-equivariant diffeomorphisms of $G \times W$. The same proof as for $\mathrm{M}(W, G)$ works in proving $H^i(Q; \mathcal{C}(W, \mathcal{Z}(G))) = 0$, $i = 1, 2$, in the case of G completely solvable. We use this fact by simply referring to the "smooth Seifert fiber space construction".

If W is a complex manifold and Q acts holomorphically, then one can also ask whether the construction can be done holomorphically on $\mathbb{C}^k \times W$. There are two types of obstructions. It could happen that the necessary $\mathbb{C}^k$-bundle over W, while trivial as a smooth bundle, may not be trivial as a holomorphic bundle. A more serious matter is that $H^2(Q; \mathcal{H}(W, \mathbb{C}^k))$, where $\mathcal{H}(W, \mathbb{C}^k)$ denotes holomorphic maps, does not necessarily vanish and so not every smooth realization has a holomorphic one. Even when the first difficulty does not arise, the latter one may still persist. The solution to these problems and the corresponding theory is carefully worked out in [13]. One particular feature of the extended theory is that the uniformizing space need no longer be a product $G \times W$ but may be any principal G-bundle over W. See [41] for Seifert fibered spaces modelled on principal bundles and 6.2 for an illustration of the holomorphic theory.

Another general approach to Seifert fiber spaces is due to Holmann [21]. His procedure extends the classical fiber bundle methods.

5. Applications

5.1. *Introduction*

The Seifert Construction, which is a special embedding, $\theta : \pi \to \mathrm{TOP}_G(G \times W)$, of the group π into $\mathrm{TOP}_G(G \times W)$ such that π acts properly on $G \times W$, preserves some of the properties of both G and W on $\theta(\pi) \setminus (G \times W)$. Furthermore, the action of π on $G \times W$ "twists" the topology and geometry of G and W to create the orbit space $\theta(\pi) \setminus (G \times W)$ in the same way that the group structures of Γ and Q "twist" to create the group π. In other words, this algebraic twisting of π makes the geometric twisting of the "bundle with singularities"

$$\Gamma \setminus G \to \theta(\pi) \setminus (G \times W) \to Q \setminus W,$$

where the homogeneous space $\Gamma \setminus G$ is a typical fiber. In the several applications, we have selected to include here this features seems especially prominent.

One of the important geometric problems that has motivated the development of Seifert fiberings is the construction of closed aspherical manifolds realizing Poincaré duality groups π of the form $1 \to \Gamma \to \pi \to Q \to 1$. The Seifert Construction enables one to find explicit aspherical manifolds $M(\pi)$ when Q acts on a contractible manifold W and Γ is a torsion free lattice in a Lie group.

For a topological manifold, the homotopy classes of self-homotopy equivalences can be regarded as algebraic data. We shall show how the Seifert Construction can be used to lift finite subgroups of homotopy classes to an action on the manifold.

The final illustration is a description of the classical 3-dimensional Seifert manifolds and how they fit into our scheme of general Seifert fiberings. This description will also be useful for the final section.

5.2. *Existence of closed smooth $K(\pi, 1)$-manifolds*

There are two difficult problems related to the title. They are:

- *Which groups can be the fundamental group of a closed aspherical manifold?* and
- *If π is the fundamental group of an aspherical manifold, can we give an actual explicit construction of an aspherical manifold for the group π?*

There are some general criteria for the first problem such as π must be finitely presented, have finite cohomological dimension and satisfy Poincaré duality in that dimension. The Seifert Construction gives answers to both questions for large classes of groups π. The idea is that if $1 \to \Gamma \to \pi \to Q \to 1$ is a torsion free extension where Γ is the fundamental group of a closed aspherical manifold and Q is a group acting properly on a contractible manifold W with compact quotient, then π should be the fundamental group of a closed aspherical manifold. We have the following

THEOREM 5.2.1. *Let Γ be a cocompact special lattice in G (see Section 4.3) and $\rho : Q \to$ $\mathrm{TOP}(W)$ be a properly discontinuous action on a contractible manifold W with compact* *quotient. If* $1 \to \Gamma \to \pi \to Q \to 1$ *is a torsion free extension of* Γ *by* Q*, then for any Seifert Construction* $\theta : \pi \to \mathrm{TOP}_G(G \times W)$,

(1) $M(\theta(\pi)) = \theta(\pi) \setminus (G \times W)$ *is a closed aspherical manifold if* Γ *is of type* (S1), (S2) *or* (S3),

(2) $M(\theta(\pi)) = \theta(\pi) \setminus ((G/K) \times W)$, *where* K *is a maximal compact subgroup of* G, *is a closed aspherical manifold if* Γ *is of type* (S4).

PROOF. We know from Theorem 4.3.2 that for each extension $1 \to \Gamma \to \pi \to Q \to 1$, there exists a homomorphism of π into $\mathrm{TOP}_G(G \times W)$ (resp., $\mathrm{TOP}_{(G,K)}(G/K \times W)$ in the case of type (S4), see [42] for this notation). We need only to check that this homomorphism is injective. Suppose Q_0 is the kernel of $\phi \times \rho : Q \to \mathrm{Out}(G) \times \mathrm{TOP}(W)$. Then Q_0 is finite since the Q action on W is properly discontinuous. Let $1 \to \Gamma \to E \to Q_0 \to 1$ be the pullback via $Q_0 \subset Q$. The group E is torsion free since π is assumed to be torsion free. The restriction $\theta|_E$ defines an action of E on $G \times W$ (resp., $G/K \times W$). Since Γ is of finite index in E, no non-trivial element of E can fix $G \times W$ (resp., $G/K \times W$). For if it did, then some power would be a non-trivial element of Γ which does not fix G (resp., G/K). Therefore, θ is injective; $\theta(\pi)$ acts properly and freely since it is torsion free. □

REMARK 5.2.2. (1) If W is a smooth contractible manifold and $\rho : Q \to \mathrm{Diff}(W)$, then the construction can be done smoothly and $M(\theta(\pi))$ is smooth.

(2) If ρ_1 and ρ_2 are rigidly related (i.e., there exists $h \in \mathrm{TOP}(W)$ for which $\rho_2 = \mu(h) \circ \rho_1$)) and Γ is characteristic in π, then $M(\theta_1(\pi))$ and $M(\theta_2(\pi))$ are homeomorphic via a Seifert automorphism. Moreover, if we fix ℓ and ρ, then the constructed $M(\theta(\pi))$ are all strictly equivalent.

(3) When $W = \{p\}$ is a point (a 0-dimensional contractible manifold), then Q must be finite for Q to act properly discontinuously, and every $\rho : Q \to \mathrm{TOP}(\{p\})$ is rigidly related. The closed aspherical manifolds constructed are infra-G-manifolds. Cf. Example 4.4.4.

(4) One important application of these constructions is that they provide model aspherical manifolds with often strong geometric properties. If one wants to study the famous conjecture that two closed aspherical manifolds with isomorphic fundamental groups are homeomorphic via the methods of controlled surgery, then the constructed aspherical Seifert manifolds are excellent model manifolds.

The above procedure can be extended for even more general extensions. As an example,

THEOREM 5.2.3. *Let* π *be a torsion-free extension of a virtually poly-*$\mathbb{Z}$ *group* Γ *by* Q, *where* Q *acts on a contractible manifold* W *properly discontinuously with compact quotient. Then there exists a closed* $K(\pi, 1)$*-manifold.*

PROOF. A torsion-free virtually poly-$\mathbb{Z}$ group Γ has a unique maximal normal nilpotent subgroup Δ, which is called the *discrete nilradical* of Γ. See [3]. Then the quotient Γ/Δ is virtually free Abelian of finite rank. Furthermore, since Δ is a characteristic subgroup of Γ, it is normal in π. Consider the commuting diagram with exact rows and columns:

$$
\begin{array}{ccccccccc}
 & & 1 & & 1 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
 & & \Delta & \xrightarrow{=} & \Delta & & & & \\
 & & \downarrow & & \downarrow & & & & \\
1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow{=} & & \\
1 & \longrightarrow & \Gamma/\Delta & \longrightarrow & \pi/\Delta & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & & & \\
 & & 1 & & 1 & & & &
\end{array}
$$

Since Γ/Δ is virtually free Abelian of finite rank (say of s), it contains a characteristic subgroup $\mathbb{Z}^s$. Let $Q' = (\pi/\Delta)/\mathbb{Z}^s$. Then the natural projection $Q' \to Q$ has a finite kernel. Therefore, if we let Q' act on W via Q, the action will still be properly discontinuous.

One can do a Seifert fiber space construction with the exact sequence

$$1 \longrightarrow \mathbb{Z}^s \longrightarrow \pi/\Delta \longrightarrow Q' \longrightarrow 1$$

which yields a properly discontinuous action of π/Δ on $\mathbb{R}^s \times W$ with compact quotient. Using this action of π/Δ on $\mathbb{R}^s \times W$, one does a Seifert fiber space construction with the exact sequence

$$1 \longrightarrow \Delta \longrightarrow \pi \longrightarrow \pi/\Delta \longrightarrow 1.$$

This gives rise to a properly discontinuous action of π on $N \times (\mathbb{R}^s \times W)$, where N is the unique simply connected nilpotent Lie group containing Δ as a lattice, with compact quotient.

If the space W is smooth, and the action of Q on W is smooth, both constructions can be done smoothly so that the properly discontinuous action of π on $N \times (\mathbb{R}^s \times W)$ is smooth.

In any case, since the group π is torsion free, the resulting action of π on $N \times (\mathbb{R}^s \times W)$ is free. Consequently, we get a closed $K(\pi, 1)$-manifold

$$M = \pi \setminus \left(N \times \mathbb{R}^s \times W\right).$$

It has a Seifert fiber structure

$$F \longrightarrow M \longrightarrow Q \setminus W,$$

where the typical fiber F itself has a Seifert fiber structure

$$\Delta \setminus N \longrightarrow F \longrightarrow T^s = \mathbb{Z}^s \setminus \mathbb{R}^s.$$

In fact, since the action of the characteristic subgroup $\mathbb{Z}^s$ on $\mathbb{R}^s$ is free, F is a genuine fiber bundle, with fiber a nilmanifold $\Delta \setminus N$ over the base torus T^s. □

The space W does not have to be aspherical. As far as the action of Q is properly discontinuous, the construction works. The resulting action of π is free if and only if the pre-image of Q_w (the stabilizer of the Q action at each $w \in W$) in π is torsion free. In this case, the space $\pi \setminus (G \times W)$ will not be aspherical. See 4.3.2.

5.3. *When is θ injective?*

Let

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle \ell} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \varphi\times\rho} & & \\ 1 & \longrightarrow & \mathrm{M}(W,G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \rtimes \mathrm{TOP}(W) & \longrightarrow & 1 \end{array}$$

be a Seifert construction. The argument as mentioned just above also shows that the action of π on a general $G \times W$ will be free (and hence θ is injective) if the pre-image of Q_w (the stabilizer of the Q action at each $w \in W$) in π is torsion free. Of course θ may be injective, that is, π acts effectively on $G \times W$, without necessarily acting freely.

We determine when the homomorphism $\theta : \pi \to \mathrm{TOP}_G(G \times W)$ has trivial kernel for special lattices Γ in G of types (S1), (S2) or (S3) in Section 4.3.1.

PROPOSITION 5.3.1. *θ is injective and $\theta(\pi) \cap \ell(G) = \theta(\Gamma)$ if and only if $\varphi \times \rho$ is injective.*

PROOF. If $\varphi \times \rho$ is injective, then clearly θ is injective and $\theta(\pi) \cap \ell(G) = \theta(\Gamma)$. Now assume that θ is injective. Let K be the kernel of $\varphi \times \rho$, and $1 \to \Gamma \to E \to K \to 1$ the pullback via $K \subset Q$. Since K acts trivially on W, it is finite and E must map injectively into $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) = \ell(G) \times_{\mathcal{Z}(G)} \mathrm{M}(W, G)$. The group K maps injectively into $\mathrm{M}(W, G)/\mathcal{Z}(G)$ by the natural projection $\ell(G) \times_{\mathcal{Z}(G)} \mathrm{M}(W, G) \to \mathrm{M}(W, G)/\mathcal{Z}(G)$, because we hypothesized that $\theta(\pi) \cap \ell(G) = \theta(\Gamma)$. We complete the argument by showing $\mathrm{M}(W, G)/\mathcal{Z}(G)$ has no non-trivial elements of finite order. This follows from the fact that the exponential map $\mathcal{G} \to G$ is bijective for our special G. Therefore the equation $x^n = a$ in G has a unique solution for any $n \in \mathbb{Z}$ and every $a \in G$. So if $f \in \mathrm{M}(W, G)$ represents an element of K in $\mathrm{M}(W, G)/\mathcal{Z}(G)$, then $f^n \in \mathcal{Z}(G)$, where n is the order of f. Therefore, $f^n(w) = a$, for some $a \in \mathcal{Z}(G)$, and all $w \in W$. But there exists a unique $x \in G$ such that $f(w) = x$. Now, $\mathcal{Z}(G)$ is $\mathbb{R}^k$ for some k, and each $a \in \mathcal{Z}(G)$ has a unique nth root in $\mathcal{Z}(G)$, $x \in \mathcal{Z}(G)$. Since f has order n, n must be 1 and $\Gamma = E$. □

5.4. *Rigidity of infra-homogeneous spaces*

Recall from 2.9 the following definitions. Let G be a connected Lie group. A quotient of G by a lattice Γ is called a *homogeneous space*. More generally, let $\mathrm{Aff}(G) = G \rtimes \mathrm{Aut}(G)$ act on G by

$$(a, \alpha) \cdot x = a \cdot \alpha(x).$$

An *infra-homogeneous space* is a quotient of G by a group $\pi \subset \mathrm{Aff}(G)$, acting properly discontinuously and freely, such that $\Gamma = \pi \cap G$ is a uniform lattice of G and Γ has a finite index in π. Therefore, an infra-homogeneous space is finitely covered by a homogeneous space. For example, a flat Riemannian manifold may be called an "infra-torus". To emphasize G, sometimes we use the term *infra-G-manifold* for an infra-homogeneous space.

Now we explain a little more detail for the claims made in Example 4.4.4. Let Δ_i be a special lattice of type (S1), (S2) or (S3). That is,

$$1 \longrightarrow \Delta_i \longrightarrow \pi_i \longrightarrow Q_i \longrightarrow 1$$

is an extension, $\theta_i : \pi_i \to \mathrm{Aff}(G_i)$, Q_i finite, $\rho_i : Q_i \to \mathrm{Out}(G_i)$ is injective, and hence by Proposition 5.3.1, $\Delta_i = \theta_i(\pi_i) \cap \ell(G_i)$, for $i = 1, 2$. Suppose $h : \pi_1 \to \pi_2$ is an isomorphism. Then, we claim, Δ_1 is mapped isomorphically onto Δ_2 and hence Q_1 onto Q_2. This isomorphism extends to an isomorphism of G_1 onto G_2. Thus we may think of

$$\theta_i : \pi \to \mathrm{Aff}(G) = \mathrm{TOP}_G\big(G \times \{p\}\big)$$

such that $\theta_i(\pi) \cap G = \Delta_i$, $i = 1, 2$. Now because $\rho_i : Q_i \to \{p\}$ is trivial, the embeddings are conjugate by an element of $\mathrm{Aff}(G)$.

It remains to verify that h maps Δ_1 isomorphically onto Δ_2. It is well known that Δ_i is characteristic in the Abelian and nilpotent cases. For a proof of Δ_i being characteristic in the completely solvable case, we refer the reader to [35, Proposition 4.5]. Therefore h maps Δ_1 isomorphically onto Δ_2.

Instead of appealing to the rigidity theorems for the Seifert Construction as we did above, we shall instead obtain these same extensions of the classical Bieberbach theorems directly. In the next theorem, the calculations involved resemble the proof of uniqueness of the Seifert Construction when W is a point, and show how closely linked the Seifert Construction is to the Bieberbach theorems.

THEOREM 5.4.1. *Suppose G has* ULIEP (*see Section* 4.3.1), *and* $H^1(\Psi; G)$ (non-Abelian cohomology) *vanishes for every finite subgroup $\Psi \subset \mathrm{Aut}(G)$. Let $\pi, \pi' \in \mathrm{Aff}(G)$ be finite extensions of lattices in G. Then every isomorphism $\theta : \pi \to \pi'$ is a conjugation by an element of* $\mathrm{Aff}(G)$.

PROOF. Let $\Gamma = G \cap \pi$, $\Gamma' = G \cap \pi'$ be the pure translations. Let $\Lambda = \Gamma \cap \theta^{-1}(\Gamma')$. Then $\theta|_\Lambda : \Lambda \to \theta(\Lambda)$ is an isomorphism of lattices of G. By ULIEP, $\theta|_\Lambda$ extends uniquely to an automorphism $C : G \to G$. Thus, $\theta|_\Lambda = C|_\Lambda$, and hence, $\theta(z, I) = (Cz, I)$ for all $(z, I) \in \Lambda$. Let us denote the composite homomorphism

$$\pi \stackrel{\theta}{\longrightarrow} \pi' \hookrightarrow \mathrm{Aff}(G) \to \mathrm{Aut}(G)$$

by $\bar{\theta}$, where $\mathrm{Aff}(G) = G \rtimes \mathrm{Aut}(G) \to \mathrm{Aut}(G)$ is the projection. Let $\Psi = \pi/\Lambda$. Then clearly, $\bar{\theta}$ factors through Ψ. Define a map $\lambda : \pi \to G$ by

$$\theta(w, K) = \big(Cw \cdot \lambda(w, K), \bar{\theta}(w, K)\big). \tag{1}$$

It is easy to see that $\lambda(zw, K) = \lambda(w, K)$ for all $z \in \Lambda$. Therefore, λ is really a map

$$\lambda : \Psi \to G.$$

For any $(z, I) \in \Lambda$ and $(w, K) \in \pi$, apply θ to both sides of $(w, K)(z, I)(w, K)^{-1} = (w \cdot Kz \cdot w^{-1}, I)$ to get

$$Cw \cdot \lambda(w, K) \cdot \bar{\theta}(w, K)(Cz) \cdot \lambda(w, K)^{-1} \cdot (Cw)^{-1} = \theta\big(w \cdot Kz \cdot w^{-1}\big).$$

However, $w \cdot Kz \cdot w^{-1} \in \Lambda$ since Λ is normal in π, and the right-hand side equals to $C(w \cdot Kz \cdot w^{-1}) = Cw \cdot CKz \cdot (Cw)^{-1}$ since $C : G \to G$ is an automorphism. From this, we have

$$\bar{\theta}(w, K)(Cz) = \lambda(w, K)^{-1} \cdot CKz \cdot \lambda(w, K). \tag{2}$$

This is true for all $z \in \Gamma$. Note that $\bar{\theta}(w, K)$, C and K are automorphisms of the Lie group G. By ULIEP of G, *the equality* (2) *holds true for all* $z \in G$.

We claim that, with the Ψ-structure on G via $\bar{\theta} : \Psi \to \mathrm{Aut}(G)$, $\lambda \in Z^1(\Psi; G)$; i.e., $\lambda : \Psi \to G$ is a crossed homomorphism. We shall show

$$\lambda\big((w, K) \cdot (w', K')\big) = \lambda(w, K) \cdot \bar{\theta}(w, K)\big(\lambda(w', K')\big)$$

for all $(w, K), (w', K') \in \pi$. (Note that we are using the elements of π to denote the elements of Ψ.) Apply θ to both sides of $(w, K)(w', K') = (w \cdot Kw', KK')$ to get $Cw \cdot \lambda(w, K) \cdot \bar{\theta}(w, K)[Cw' \cdot \lambda(w', K')] = C(w \cdot Kw') \cdot \lambda((w, K)(w', K'))$. From this, it follows that

$$\lambda\big((w, K)(w', K')\big) = (CKw')^{-1} \cdot \lambda(w, K) \cdot \bar{\theta}(w, K)(Cw') \cdot \bar{\theta}(w, K)\lambda(w', K').$$

By (2), we have $\bar{\theta}(w, K)(Cw') = \lambda(w, K)^{-1} \cdot CKw' \cdot \lambda(w, K)$ so that $\lambda((w, K) \cdot (w', K')) = \lambda(w, K) \cdot \bar{\theta}(w, K)\lambda(w', K')$. This shows that λ is a crossed homomorphism.

By the assumption, $H^1(\Psi; G)$ vanishes. This means that any crossed homomorphism is "principal". In other words, there exists $d \in G$ such that

$$\lambda(w, K) = d \cdot \bar{\theta}(w, K)\big(d^{-1}\big). \tag{3}$$

Let $D = \mu(d^{-1}) \circ C$. We check that θ is the conjugation by $(d, D) = (d, \mu(d^{-1}) \circ C) \in \mathrm{Aff}(G)$. Using (1), (2) and (3), one can show $\bar{\theta}(w, K) \circ \mu(d^{-1}) \circ C = \mu(d^{-1}) \circ C \circ K$. Thus, for any $(w, K) \in \pi$,

$$\begin{aligned}
&\theta(w, K) \cdot (d, D) \\
&\quad = \big(Cw \cdot \lambda(w, K), \bar{\theta}(w, K)\big) \cdot \big(d, \mu\big(d^{-1}\big) \circ C\big) \\
&\quad = \big(Cw \cdot \lambda(w, K) \cdot \bar{\theta}(w, K)(d), \bar{\theta}(w, K) \circ \mu\big(d^{-1}\big) \circ C\big) \\
&\quad = \big(Cw \cdot d \cdot \bar{\theta}(w, K)\big(d^{-1}\big) \cdot \bar{\theta}(w, K)(d), \bar{\theta}(w, K) \circ \mu\big(d^{-1}\big) \circ C\big)
\end{aligned}$$

$$= \big(Cw \cdot d, \mu(d^{-1}) \circ C \circ K\big)$$
$$= (d, D) \cdot (w, K).$$

This finishes the proof of our claim. □

COROLLARY 5.4.2. *Suppose G has* ULIEP, *and $H^1(\Psi; G)$* (non-Abelian cohomology) *vanishes for every finite subgroup $\Psi \subset \mathrm{Aut}(G)$. Then homotopy equivalent infra-G-manifolds are affinely diffeomorphic.*

PROOF. Let $M = \pi \setminus G$, $M' = \pi' \setminus G'$ be infra-G-manifolds. A homotopy equivalence $M \to M'$ induces an isomorphism $\theta : \pi \to \pi'$. Since θ maps some subgroup of $G \cap \pi$ into a lattice of G', there is an isomorphism of G onto G'. Using this isomorphism, we identify G with G'. Now apply Theorem 5.4.1 to $\theta : \pi \to \pi'$ to find an element $h = (d, D) \in \mathrm{Aff}(G)$ which conjugates π onto π'. This gives a weakly equivariant map

$$\big(h, \mu(h)\big) : (G, \pi) \to (G, \pi')$$

and h gives rise to an affine diffeomorphism $M \to M'$, which is homotopic to the original map. □

COROLLARY 5.4.3. (1) (Bieberbach) *Homotopy equivalent flat manifolds are affinely diffeomorphic.*

(2) [39] *Homotopy equivalent infra-nilmanifolds are affinely diffeomorphic.*

(3) *Homotopy equivalent infra-solvmanifolds of type* (R) *are affinely diffeomorphic.*

Notice, in particular, that the isomorphism θ maps the lattice $\Gamma = G \cap \pi$ onto $\Gamma' = G \cap \pi'$. This is also true topologically on the orbifold level.

PROOF. What is needed for this conclusion is the element $h = (d, D) \in \mathrm{Aff}(G)$ which conjugates π onto π'. We have this element immediately from the rigidity of the Seifert Construction for lattices of type (S1), (S2) and (S3). If we wish to avoid the Seifert Construction, then we need to verify that $H^1(\Psi; G)$ vanishes for each finite subgroup $\Psi \subset \mathrm{Aut}(G)$. This is accomplished in [39] and [35], and well known for G of type (S1). □

5.5. *Lifting problem for homotopy classes of self-homotopy equivalences*

Let M be an admissible space (see Section 2.5) and $\mathcal{E}(M)$ be the H-space of homotopy equivalences of M into itself. Any $f \in \mathcal{E}(M)$ induces an isomorphism $f_* : \pi_1(M, x) \to \pi_1(M, f(x))$. By choosing a path ω from x to $f(x)$, we have an automorphism f_*^{ω} of $\pi_1(M, x)$, defined by $f_*^{\omega}([\tau]) = [\omega^{-1} \cdot (f \circ \tau) \cdot \omega]$. A different choice of ω alters f_*^{ω} only by an inner automorphism. Therefore, we obtain a homomorphism

$$\gamma : \mathcal{E}(M) \to \mathrm{Out}(\pi),$$

$\pi = \pi_1(M, x)$. If M is a $K(\pi, 1)$ space, then $\mathcal{E}_0(M)$ is the kernel of γ so that γ factors through $\pi_0(\mathcal{E}(M)) = \mathcal{E}(M)/\mathcal{E}_0(M)$, where $\mathcal{E}_0(M)$ is the self homotopy equivalences homotopic to the identity. Moreover, γ is onto since every automorphism of π can be realized by a self homotopy equivalence of M.

A homomorphism $\varphi: F \to \mathrm{Out}(\pi) \cong \pi_0(\mathcal{E}(M))$ is called an *abstract kernel*. An injective abstract kernel is the same as a subgroup of homotopy classes of self-homotopy equivalences of M. A *lifting* of φ as a group of homeomorphisms is a homomorphism $\hat{\varphi}: F \to \mathrm{TOP}(M)$ which makes

$$\begin{array}{ccccc} F & & \xrightarrow{=} & & F \\ \downarrow{\scriptstyle\hat{\varphi}} & & & & \downarrow{\scriptstyle\varphi} \\ \mathrm{TOP}(M) & \longrightarrow & \mathcal{E}(M) & \longrightarrow & \pi_0(\mathcal{E}(M)) \end{array}$$

commutative.

Suppose F acts on M. Let F^* be the group of all liftings of elements of F to the universal covering $\widetilde{M}$. Then

$$1 \to \pi \to F^* \to F \to 1$$

is exact, and is called the *lifting sequence* of the action (F, M). Furthermore, for an aspherical M, if (F, M) is effective, then *the centralizer of* π *in* F^*, $C_{F^*}(\pi)$, *is torsion-free* (such an extension was called *admissible* in [36]); and F^* lies in $\mathrm{TOP}(\widetilde{M})$.

Thus we have a necessary condition for the existence of a lifting of an abstract kernel $\varphi: F \to \mathrm{Out}(\pi)$, as an (effective, resp.) group action: the existence of an (admissible, resp.) group extension of π by F realizing the abstract kernel. For finite groups, this necessary condition is also sufficient for some tractable manifolds.

No examples of closed aspherical manifolds are yet known where the existence of a group extension does not also yield a group action.

DEFINITION 5.5.1. Let Q act properly on an admissible space W (see Section 2.5) and B be the quotient $Q \setminus W$. Suppose for each extension $1 \to Q \to E \to F \to 1$ by a finite group F, the action of Q extends to a proper action of E on W. Then we say that the Q action on W is *finitely extendable*. In particular, then F acts on B preserving the orbit structure.

If Γ is normal in π, let $\mathrm{Aut}(\pi, \Gamma)$ denote the automorphisms of π that leave Γ invariant. Since $\mathrm{Inn}(\pi)$ leaves Γ invariant, we can put $\mathrm{Aut}(\pi, \Gamma)/\mathrm{Inn}(\pi) = \mathrm{Out}(\pi, \Gamma)$. It is a subgroup of $\mathrm{Out}(\pi)$. See notation in Section 1.

Let $M = \pi \setminus \widetilde{M}$ be an orbifold (with the orbifold group π). Call a finite abstract kernel $F \to \mathrm{Out}(\pi)$ *geometrically realizable* if it can be realized as an action of F on M.

We are interested in realizing a finite abstract kernel $F \to \mathrm{Out}(\pi)$ as a group action on a Seifert fiber space M with a typical fiber $\Gamma \setminus G$. In particular, we want the F action to be fiber-preserving maps. This means that, on the group level, the extension must leave the lattice Γ invariant. In other words, we consider only those abstract kernels which have images in $\mathrm{Out}(\pi, \Gamma)$.

THEOREM 5.5.2. *Let Γ be a lattice in the special Lie group G (see Section 4.3.1), and $\rho: Q \to \mathrm{TOP}(W)$ be a proper action on W which is finitely extendable. Let $1 \to \Gamma \to \pi \to Q \to 1$ be an extension and $\theta: \pi \to \mathrm{TOP}_G(G \times W)$ be any homomorphism. Then each abstract kernel $\varphi: F \to \mathrm{Out}(\pi, \Gamma)$ of a finite group F can be geometrically realized as a group of Seifert automorphisms on $M(\theta(\pi)) = \theta(\pi) \setminus (G \times W)$ if and only if the abstract kernel φ admits some extension.*

PROOF. Let $1 \to \pi \to E \to F \to 1$ be an extension realizing the abstract kernel φ. Consider the induced extension

$$1 \longrightarrow Q \longrightarrow E/\Gamma \longrightarrow F \longrightarrow 1.$$

Since ρ is finitely extendable, there exists $\rho': E/\Gamma \to \mathrm{TOP}(W)$ extending $\rho: Q \to \mathrm{TOP}(W)$. We have the commutative diagram with exact columns and rows:

$$\begin{array}{ccccccccc}
 & & & & 1 & & 1 & & \\
 & & & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \Gamma & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \Gamma & \longrightarrow & E & \longrightarrow & E/\Gamma & \longrightarrow & 1 \\
 & & & & \downarrow & & \downarrow & & \\
 & & & & F & \xrightarrow{=} & F & & \\
 & & & & \downarrow & & \downarrow & & \\
 & & & & 1 & & 1 & &
\end{array}$$

Again by the existence theorem for special lattices, Theorem 4.3.2, there exists $\theta': E \to \mathrm{TOP}_G(G \times W)$, where $\theta'|_\Gamma = \theta|_\Gamma = i: \Gamma \hookrightarrow G$, and $\rho'|_Q = \rho$. Put $\theta'|_\pi = \theta'$. Of course, θ' may be different from θ, but as θ and θ' agree on Γ and Q, we can apply Theorem 4.3.2(2) to conjugate $\mathrm{TOP}_G(G \times W)$ by an element of $\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G)$ which carries $\theta'|_\pi$ to θ so that the new homomorphism $\theta': E \to \mathrm{TOP}_G(G \times W)$ is an extension of $\theta: \pi \to \mathrm{TOP}_G(G \times W)$. This yields an action of F on $\theta(\pi) \setminus (G \times W)$ as a group of Seifert automorphism as desired. □

Since we used the uniqueness of the Seifert Construction, we use the modified Seifert Construction modelled on $G/K \times W$ instead of $G \times W$ in the semi-simple case.

COROLLARY 5.5.3. *Let $M(\pi)$ be an infra-G-manifold with G special, and with fundamental group π. Suppose $\varphi: F \to \mathrm{Out}(\pi)$ is an abstract kernel. Then there exists a geo-*

metric realization of F acting on $M(\pi)$ by affine diffeomorphisms if and only if there is an extension E of π by F which realizes this abstract kernel.

PROOF. The infra-G-manifold $M(\pi)$ is modelled on $G \times \{p\}$ ($p =$ point) ($G/K \times \{p\}$ for a convenient form of G in the semi-simple case) and $\mathrm{TOP}_G(G \times \{p\}) = \mathrm{Aff}(G)$ (resp. $\overline{\mathrm{Aff}}(G, K)$), see [42] for this notation. Trivially, every $Q \to \mathrm{TOP}(\{p\})$ extends to $E/\Gamma \to \mathrm{TOP}(\{p\})$. The above theorem then immediately applies, and F acts on $M(\pi)$ by Seifert automorphisms which are affine diffeomorphisms. □

REMARK 5.5.4. (1) Since we may introduce a metric structure in the corollary from a left invariant metric on G (see Section 2.9), $M(\pi)$ has the structure of a flat, almost flat, Riemannian infra-solvmanifold or a locally symmetric spaces. We may also further conjugate $\theta(\pi)$ in $\mathrm{Aff}(G)$ so that F now acts on the conjugated manifold by isometries preserving the flat, etc., structures.

(2) The proper action of π in the theorem is not necessarily free nor effective. Thus $M(\pi)$ could very well be a Seifert orbifold. The corollary then works for such orbifolds, i.e., infra-G-spaces. In the Euclidean case, $M(\pi)$ would then be a Euclidean "crystal" and π a Euclidean crystallographic group. In Theorem 5.5.2, F sends fibers to fibers.

For more about the realizations up to strict equivalences, and finding examples where F does not lift because there are no extensions realizing the abstract kernels, the reader is referred to [22,26,32,33,36,37,42,53,59,64] and [52].

5.6. *Seifert fiberings with codimension 2 fibers*

5.6.1. There is a class of Seifert fiberings which are very close generalizations of the classical Seifert 3-manifolds. Let the discrete group Q act effectively and properly on the topological plane $\mathbb{R}^2$. Then Q can be conjugated in $\mathrm{TOP}(\mathbb{R}^2)$ to be a group acting as isometries on the hyperbolic plane or the Euclidean plane. In the first case, call Q hyperbolic, and in the latter Euclidean. If Q preserves orientation, then Q can be conjugated to act as holomorphic actions on $\mathbb{C}$ or the unit disk D. This latter fact, using standard techniques in transformation group theory, uniformization and topology of 2-manifolds, is not very difficult to prove. (Certainly, if Q acts simplicially, the arguments are not hard.) When the orbit space $Q \setminus \mathbb{R}^2$ is compact, then a theorem, essentially due to Nielsen, says that any two Q actions are topologically conjugate. If the actions are assumed smooth, then they are smoothly conjugate. In the non-compact case, there are similar results but they can be very complicated if Q is not finitely presentable. So for our next example, we restrict to Q for which $Q \setminus \mathbb{R}^2$ is compact. Since $Q \setminus \mathbb{R}^2$ will have the structure of an orbifold, Q will either be isomorphic to a Euclidean crystallographic group or a cocompact hyperbolic group. In the Euclidean case, these are the 17 wall-paper groups. They are centerless except for $Q = \mathbb{Z} \times \mathbb{Z}$ or $\mathbb{Z} \rtimes \mathbb{Z}$ (the fundamental group of the Klein bottle). In the hyperbolic case, all the groups are not solvable, and have no non-trivial normal Abelian subgroups.

5.6.2. An important class of closed 3-manifolds are those closed 3-manifolds whose fundamental group contains a normal subgroup $\mathbb{Z}$. This class contains all the classical Seifert 3-manifolds with infinite fundamental group as well as some other non-orientable manifolds such as closed 3-manifolds admitting a circle action without fixed points but having

2-dimensional submanifolds whose stabilizer is $\mathbb{Z}_2$. The orientable ones, however, coincide with the classical orientable Seifert 3-manifolds with infinite fundamental group as defined by Seifert in [60]. Of course, these statements are to be taken modulo the Poincaré Conjecture, for if there were a fake 3-sphere Σ, then $M \sharp \Sigma$, where M is a Seifert manifold, would not be a Seifert manifold but would still have a normal $\mathbb{Z}$ in its fundamental group. In any case, a closed 3-manifold whose fundamental group contains a normal $\mathbb{Z}$ subgroup, has, modulo the Poincaré Conjecture, an orientable cover admitting an injective S^1-action. Consequently, they all have a universal cover that splits into $\mathbb{R} \times \mathbb{R}^2$ or $\mathbb{R} \times S^2$, where the $\mathbb{R}$ descends to the fibers associated with the normal $\mathbb{Z}$. In the $\mathbb{R} \times S^2$ case, the manifolds have an orientable cover isomorphic to $S^1 \times S^2$. Thus, their fundamental groups are just finite extensions of $\mathbb{Z}$. We are interested in those that are not finite extensions of $\mathbb{Z}$. They are all aspherical since they are covered by $\mathbb{R}^3$.

Therefore, for each extension, $1 \to \mathbb{Z} \to \pi \to Q \to 1$, where $\rho : Q \to \mathrm{TOP}(\mathbb{R}^2)$ is an effective proper action on $\mathbb{R}^2$, we may construct model Seifert manifolds. In fact, as ρ is topologically rigid, we can assume that after conjugating in $\mathrm{TOP}(\mathbb{R}^2)$, ρ maps Q into either $\mathrm{Isom}(\mathbb{E}^2) = \mathbb{R}^2 \rtimes \mathrm{O}(2)$ or $\mathrm{Isom}(\mathbb{H}) = \mathrm{PSL}(2, \mathbb{R}) \rtimes \mathbb{Z}_2$, the isometries of the Euclidean plane or the hyperbolic plane. In this case, as we shall see in Section 6.3, the constructed Seifert manifolds possess a geometric structure inherited from a subgroup $\mathcal{U}$ of $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{R}^2)$ into which $\pi_1(M)$ is embedded. The group π acts freely on $\mathbb{R} \times \mathbb{R}^2$ if and only if π is torsion free. Otherwise, the map $\mathbb{R} \times \mathbb{R}^2 \to \theta(\pi) \setminus (\mathbb{R} \times \mathbb{R}^2)$ is a non-trivially branched covering and the underlying topological space of the 3-dimensional orbifold, which is still a 3-manifold with possible boundary, may not even be aspherical (see the next section). To obtain all possible orbifolds, we have to enlarge Q to Q', where Q is of index 2 in Q' and the $\mathbb{Z}_2$ in Q' does not act effectively (it injects into $\mathrm{Aut}(\mathbb{Z})$). That is, $\varphi \times \rho : Q' \to \mathrm{Aut}(\mathbb{Z}) \times \mathrm{TOP}(\mathbb{R}) \to \mathrm{Aut}(\mathbb{R}) \times \mathrm{TOP}(\mathbb{R}^2)$ will be injective.

Obviously the existence of the Seifert Construction for every torsion free extension $1 \to \mathbb{Z} \to \pi \to Q \to 1$ produces all the known aspherical 3-manifolds with a normal $\mathbb{Z}$ in their fundamental group. Furthermore, rigidity of the construction, since $\mathbb{Z}$ will almost always be characteristic, says that any two constructed manifolds with isomorphic fundamental groups are diffeomorphic. In fact, except for several exceptions (less than 10), the isomorphisms must be given by a Seifert automorphism. (The exceptions can occur for the 3-torus, for example, where the $\mathbb{Z}$ of the fiber is not the only possible normal $\mathbb{Z}$.) That no other manifolds are possible except those given by the Seifert Construction follows from results of Waldhausen in the Haken case. For non-Haken manifolds, the result is more recent and is due to Scott, Casson and others.

The rigidity result is also true in the orbifold case. For example, if Q is *orientation-preserving* and $Q \to \mathrm{Aut}(\mathbb{Z})$ is trivial, then the $\mathbb{R}$ action descends to an S^1-action on $\theta(\pi) \setminus (\mathbb{R} \times \mathbb{R}^2)$. The group $\mathbb{Z}$ is the maximal normal Abelian subgroup of π in case Q is hyperbolic since Q has no normal Abelian subgroups. In the Euclidean crystallographic case where $Q \neq \mathbb{Z}^2$, Q is centerless and so $\mathbb{Z}$ will be the entire center of π and so is characteristic. For $Q = \mathbb{Z}^2$, the π are torsion free and are distinguished by $H_1(\pi; \mathbb{Z}) = \mathbb{Z}^2 + \mathbb{Z}_b$. So, except for 3-torus, *the rigidity theorem for the Seifert Construction implies two such oriented orbifolds with isomorphic orbifold groups will be diffeomorphic via an S^1-equivariant orientation-preserving homeomorphism.*

THEOREM 5.6.3 (Higher-dimensional fibers). *Let Q be hyperbolic and act properly on $\mathbb{R}^2$ with compact quotient. Let Γ be a special lattice of type* (S1), (S2) *or* (S3) (*see Section* 4.3.1), *and*

$$1 \longrightarrow \Gamma \longrightarrow \pi \longrightarrow Q \longrightarrow 1$$

be an extension. Then there exists a homomorphism $\theta : \pi \to \mathrm{TOP}_G(G \times \mathbb{R}^2)$ unique up to conjugation. π acts freely if and only if π is torsion free and, consequently, $M(\theta(\pi))$ will then be aspherical. In general, $M(\theta_1(\pi_1))$ is homeomorphic via a Seifert automorphism to $M(\theta_2(\pi_2))$ if and only if π_1 is isomorphic to π_2.

PROOF. We know θ exists by the existence theorem for special lattices. Since Q has no non-trivial normal solvable subgroups, and G is solvable, $\pi \cap G = \Gamma$. Therefore, Γ is the maximal normal solvable subgroup in π and consequently is characteristic. Since $\rho : Q \to \mathrm{TOP}(\mathbb{R}^2)$ is rigid, the Seifert Construction is rigid and the theorem follows. (Cf. [13, Section 12].) □

Nicas and Stark in [47], using controlled surgery, have shown that any closed aspherical topological manifold whose fundamental group is a central extension of Z^s by a 2-dimensional orientation preserving cocompact hyperbolic Q is homeomorphic to the corresponding Seifert manifold constructed in Theorem 5.6.3, provided the dimension of the manifold is greater than 4.

COROLLARY 5.6.4. *Let $\varphi : F \to \mathrm{Out}(\pi)$ be an abstract kernel, where F is finite and π is as in the above theorem. Then φ has a geometric realization of an action of F as a group of Seifert automorphism on $M(\theta(\pi))$ if and only if there exists some extension E of π which realizes the abstract kernel.*

PROOF. The extension E induces a short exact sequence $1 \to Q \to E/\Gamma \to F \to 1$. S. Kerckhoff [23] has shown, in his solution to the Nielsen Realization Problem, that a finite extension of a hyperbolic Q with an injective abstract kernel is again a hyperbolic group. If K denotes the kernel of the abstract kernel $F \to \mathrm{Out}(Q)$, then $(E/\Gamma)/K$ is a hyperbolic group containing a topological conjugate, in $\mathrm{TOP}(\mathbb{R}^2)$, of Q. Then E/Γ, via $E/\Gamma \to (E/\Gamma)/K$, acts properly on $\mathbb{H}$ and (up to conjugation in $\mathrm{TOP}(\mathbb{R}^2)$) extends the Q action. Therefore Theorem 5.5.2 applies. □

5.7. *The classical Seifert 3-dimensional manifolds*

In a classical paper [60] in 1933, H. Seifert described a class of 3-dimensional manifolds that have turned out to be fundamental for the topology of 3-manifolds. We assume the reader has some familiarity with these manifolds.

In the closed orientable (resp. non-orientable) case Seifert's manifolds whose fundamental groups are infinite, but not a finite extension of $\mathbb{Z}$, coincide with (resp. are a subset of) those closed orientable (resp. non-orientable) manifolds, $\mathcal{SF}$, that can be constructed, as in Proposition 4.3.2, from a torsion free extension, $1 \to Z \to \pi \to Q \to 1$, with Q acting properly, effectively, and cocompactly on $\mathbb{R}^2$.

It has recently been shown that if there are no fake homotopy 3-spheres, then the closed 3-manifolds, whose fundamental groups contain a normal $\mathbb{Z}$ subgroup and which are not finite extensions of $\mathbb{Z}$, coincide with $\mathcal{SF}$. See 5.6.2.

6. Reduction of the universal group

6.1. *Purpose and requirements*

In a general Seifert Construction, it may happen that $\rho: Q \to \mathrm{TOP}(W)$ has an image in a subgroup that has geometric or topological significance. This means that it is likely that the associated Seifert Constructions inherit some of these properties. Let $\mathcal{U}$ be a subgroup of $\mathrm{TOP}_G(G \times W)$ containing $\ell(G)$. Put $\widehat{\mathcal{U}} = \mathcal{U} \cap (\mathrm{M}(W, G) \rtimes \mathrm{Inn}(G))$ and $\overline{\mathcal{U}} = \mathcal{U}/\widehat{\mathcal{U}}$ so that

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \widehat{\mathcal{U}} & \longrightarrow & \mathcal{U} & \longrightarrow & \overline{\mathcal{U}} & \longrightarrow & 1 \\ & & \cap & & \cap & & \cap & & \\ 1 & \longrightarrow & \mathrm{M}(W, G) \rtimes \mathrm{Inn}(G) & \longrightarrow & \mathrm{TOP}_G(G \times W) & \longrightarrow & \mathrm{Out}(G) \times \mathrm{TOP}(W) & \longrightarrow & 1 \end{array} \tag{6.1}$$

commutes.

If the image of $\theta: \pi \to \mathrm{TOP}_G(G \times W)$ lies in $\mathcal{U}$, we say that the universal group has been reduced to $\mathcal{U}$ for π. The group $\mathcal{U}$ can then be used to study extra structures on the Seifert fiber space $\theta(\pi) \setminus (G \times W)$.

We have already seen in Section 4.4 that if W is a smooth manifold, Γ special and $\rho: Q \to \mathrm{Diff}(W)$, then any injective Seifert Construction can be done smoothly in

$$\mathrm{Diff}_G(G \times W) = \mathcal{C}(W, G) \rtimes \big(\mathrm{Aut}(G) \times \mathrm{Diff}(W)\big),$$

where $\mathrm{Diff}_G(G \times W) = \mathrm{TOP}_G(G \times W) \cap \mathrm{Diff}(G \times W)$.

Furthermore, existence, uniqueness and rigidity also hold because we may use a differential partition of unity in the proof for the vanishing of the necessary cohomology groups.

To obtain existence of an injective Seifert Construction in $\mathcal{U}$, one has to verify the conditions in Theorem 3.7.2 in $\mathcal{U}$. If $\mathcal{U}$ is much smaller than $\mathrm{TOP}_G(G \times W)$, uniqueness and rigidity are not likely to hold. The set of all homomorphisms of π into $\mathcal{U}$ which belong to the same conjugacy class in $\mathrm{TOP}_G(G \times W)$ is a deformation space of that conjugacy class. We shall illustrate these concepts by describing in some detail the deformation spaces for interesting geometric situations.

6.2. *Injective holomorphic Seifert fiberings*

We assume that $G = \mathbb{C}^k$, W is a complex manifold, $\rho: Q \to \mathrm{TOP}(W)$ has image in the holomorphic homeomorphisms of W, and $\mathcal{H}(W, \mathbb{C}^k) \subset \mathrm{M}(W, \mathbb{C}^k)$ are the holomorphic maps.

We also assume that the map $\mathbb{C}^k \times W \to W$ is holomorphically trivial and so $\mathcal{U} = \mathrm{Hol}_{\mathbb{C}^k}(\mathbb{C}^k \times W)$ then becomes

$$\mathcal{H}(W, \mathbb{C}^k) \rtimes \left(\mathrm{GL}(k, \mathbb{C}^k) \times \mathrm{Hol}(W)\right),$$

where $\mathrm{Hol}(W)$ is the group of holomorphic automorphisms of W. Existence, uniqueness and rigidity do not necessarily hold because we do not have holomorphic partitions of unity and the groups $H^i(Q; \mathcal{H}(W, \mathbb{C}^k))$ does not vanish in general.

The reader is referred to [13] where a general and comprehensive theory of holomorphic Seifert fiberings whose universal space is a holomorphic fiber bundle over W with fiber a complex torus or $\mathbb{C}^k$ is given. We shall restrict ourselves here to a special case closely related to the classical Seifert 3-manifolds.

Let $(\mathbb{C}^*, M)$ be an injective, proper, holomorphic $\mathbb{C}^*$ action on a complex 2-manifold M so that the quotient space is compact. As in the case of an injective S^1 action, an injective proper $\mathbb{C}^*$ action lifts to the covering space of M corresponding to the image of the evaluation homomorphism, and yields a splitting $(\mathbb{C}^*, \mathbb{C}^* \times W)$ where W is a simply connected complex 1-manifold (cf. Section 2.2). Therefore, W is $\mathbb{C}$, D the open unit disk, or $\mathbb{C}P_1$. We shall restrict ourselves to W being the unit disk D. The orbit space $Q \setminus W$ is a closed Riemann surface. The action of $Q = \pi_1(M)/\mathbb{Z}$ on D is holomorphic, properly discontinuous, but not necessarily free. Therefore, $M \to \mathbb{C}^* \setminus M = Q \setminus W$ is a generalization of a principal holomorphic $\mathbb{C}^*$-bundle over a Riemann surface.

From the exact sequence $1 \to \mathbb{Z} \to \mathbb{C} \xrightarrow{\exp} \mathbb{C}^* \to 0$, we obtain the exact sequence

$$0 \longrightarrow \mathbb{Z} = M(W, \mathbb{Z}) \longrightarrow \mathcal{H}(W, \mathbb{C}) \longrightarrow \mathcal{H}(W, \mathbb{C}^*) \longrightarrow 0$$

which gives rise to a long exact sequence of cohomology groups

$$\cdots \xrightarrow{\delta^{i-1}} H^i(Q; \mathbb{Z}) \longrightarrow H^i\left(Q; \mathcal{H}(W, \mathbb{C})\right) \longrightarrow H^i\left(Q; \mathcal{H}(W, \mathbb{C}^*)\right) \xrightarrow{\delta^i} \cdots.$$

The group Q acts on the unit disk D as a cocompact Fuchsian group. That is, $\rho : Q \to \mathrm{Hol}(D)$, the complex automorphisms of the unit disk. The action of Q on $\lambda \in \mathcal{H}(W, \mathbb{C})$ is given by

$${}^{\alpha}\lambda = \lambda \circ \left(\rho(\alpha)\right)^{-1}.$$

Let us compare the smooth situation with the holomorphic one. We have the following commutative diagram of exact sequences

$$\begin{array}{ccccccccc}
0 \to H^1(Q, \mathbb{Z}) & \to & H^1(Q, \mathcal{H}(W, \mathbb{C})) & \to & H^1(Q, \mathcal{H}(W, \mathbb{C}^*)) & \xrightarrow{\delta} & H^2(Q, \mathbb{Z}) & \to & H^2(Q; \mathcal{H}(W, \mathbb{C})) \\
\downarrow = & & \downarrow & & \downarrow & & \downarrow = & & \downarrow \\
\to H^1(Q, \mathbb{Z}) & \to & H^1(Q, \mathcal{C}(W, \mathbb{C})) & \to & H^1(Q, \mathcal{C}(W, \mathbb{C}^*)) & \xrightarrow{\delta} & H^2(Q, \mathbb{Z}) & \to & H^2(Q; \mathcal{C}(W, \mathbb{C}))
\end{array} \tag{6.2}$$

For the smooth case, $H^i(Q; \mathcal{C}(W, \mathbb{C})) = 0$, $i > 0$, and as we shall see, $H^2(Q; \mathcal{H}(W, \mathbb{C})) = 0$.

For each central extension $0 \to \mathbb{Z} \to \pi \to Q \to 0$ represented by $[f] \in H^2(Q;\mathbb{Z})$, we have smooth Seifert Constructions $\theta : \pi \to \mathrm{Diff}_{\mathbb{C}}(\mathbb{C} \times D) = \mathrm{Diff}_{\mathbb{C}}(\mathbb{C} \times \mathbb{R}^2)$. If we fix $i : \mathbb{Z} \to \mathbb{C}$ and $\rho : Q \to \mathrm{Diff}(D)$, the construction is unique up to strict equivalences (Section 4.4.2). We have the smooth Seifert orbifold over $Q \setminus W = Q \setminus D$ with an induced $\mathbb{C}^*$ action and therefore an S^1 action on $\theta(\pi) \setminus (\mathbb{C} \times D)$ $(\cong \mathbb{R}^1 \times \theta(\pi) \setminus (\mathbb{R}^1 \times D) = \mathbb{R}^1 \times N^3$ since $\mathbb{C}^*$ splits smoothly as $\mathbb{R}^1 \times S^1)$. The uniqueness says that for any other embedding $\theta' : \pi \to \mathrm{Diff}_{\mathbb{C}}(\mathbb{C} \times D)$, keeping i and ρ fixed, the $\mathbb{C}^*$ action on $\theta'(\pi) \setminus (\mathbb{C} \times D)$ is strictly smoothly equivalent to that on $\theta(\pi) \setminus (\mathbb{C} \times D)$.

For the same π, we have the homomorphism $H^2(Q;\mathbb{Z}) \to H^2(Q;\mathcal{H}(D,\mathbb{C}))$. The second group fortunately can be identified with the second cohomology of the sheaf of germs of holomorphic functions over $Q \setminus D$. This vanishes since $Q \setminus D$ is (complex) 1-dimensional and the sheaf is coherent (i.e., locally free). This means that $[f] \in H^2(Q;\mathbb{Z})$ maps to $0 \in H^2(Q;\mathcal{H}(W,\mathbb{C}))$. But as the groups are Abelian, this becomes exactly the identity (4.1), and we have $\theta : \pi \to \mathrm{Hol}_{\mathbb{C}}(\mathbb{C} \times W)$. Therefore, each $[f]$ has holomorphic realizations for each fixed i and ρ. Recall from Theorem 3.7.2, the set of all $\theta : \pi \to \mathrm{Hol}_{\mathbb{C}}(\mathbb{C} \times D)$ with fixed $i : \mathbb{Z} \to \mathbb{C}$ and $\rho : Q \to \mathcal{H}(D)$, up to conjugation by elements of $\mathcal{H}(D,\mathbb{C})$, is in one-one correspondence with $H^1(Q;\mathcal{H}(D,\mathbb{C}))$. (This complex vector space is the same as $H^1(V; h^0_{\mathbb{C}})$, the first cohomology of the sheaf of germs of holomorphic functions where V is treated as the analytic space $V = Q \setminus D$. That is, for each open U in V, we consider $p^{-1}(U)$ and holomorphic functions $\lambda : p^{-1}(U) \to \mathbb{C}$ such that $\lambda(\rho(\alpha)(w)) = \lambda(w)$, $w \in p^{-1}(U)$ and $p : D \to Q \setminus D$ is the projection. This defines the sheaf $h^0_{\mathbb{C}}$ over V.) This group is isomorphic to $\mathbb{C}^g$, where g is the genus of V.

THEOREM 6.2.1 ([13, §13]). *For each smooth action* $(\mathbb{C}^*, M)$ *corresponding to the unique strict conjugacy class* $\theta(\pi)$, *with* π *torsion free, there exists a complex* g*-dimensional family of strictly holomorphically inequivalent* $\mathbb{C}^*$ *actions each strictly smoothly equivalent to the smooth* $(\mathbb{C}^*, M)$.

PROOF. We may interpret $[f'] \in H^1(Q;\mathcal{H}(W,\mathbb{C}^*))$ to represent the effective holomorphic $\mathbb{C}^*$ action on $\theta(\pi) \setminus (\mathbb{C} \times D)$ up to strict $\mathbb{C}^*$ equivalence. If Q were torsion free, then Q acts freely and $Q \setminus D = V$ is a closed oriented surface without branch points. The $\mathbb{C}^*$ action is then free and proper yielding a principal holomorphic $\mathbb{C}^*$ bundle, corresponding to a complex line bundle over V. Since Q is not assumed to be torsion free, f' determines the holomorphic $\mathbb{C}^*$ action (cf. [13, §5]). Given two injective $\mathbb{C}^*$ actions $[f']$ and $[f'']$ with the same "Chern class" $\delta[f'] = \delta[f'']$, there exists a $[\lambda] \in H^1(Q;\mathcal{H}(D,\mathbb{C}))$ such that $[f''] = [f'] + [\lambda]$. Since $H^1(Q;\mathcal{H}(D,\mathbb{C}))$ is a vector space of complex dimension g, there exists a whole g-dimensional family of inequivalent holomorphic $\mathbb{C}^*$ actions starting from $[f']$ and ending with $[f'']$. □

Returning to diagram (6.2), define

$$\begin{aligned}\mathrm{Pic}(Q \setminus D) &= H^1\big(Q;\mathcal{H}(D,\mathbb{C})\big)/\mathrm{image}\big(H^1(Q;\mathbb{Z})\big)\\ &= \text{a complex } g\text{-torus, or real } 2g\text{-torus, } T^{2g}.\end{aligned}$$

The connected component of the group $H^1(Q;\mathcal{H}(D,\mathbb{C}^*))$, in the Q torsion free case, is called the *Picard group* for the line bundles over $Q \setminus D$. In our case, $H^1(Q;\mathcal{H}(D,\mathbb{C}^*))$ is isomorphic to $T^{2g} \oplus \mathbb{Z} \oplus$ finite torsion.

We obtain the exact sequence

$$0 \longrightarrow \operatorname{Pic}(Q \setminus D) \longrightarrow H^1\bigl(Q;\mathcal{H}(D,\mathbb{C}^*)\bigr) \xrightarrow{\delta} H^2(Q;\mathbb{Z}) \longrightarrow 0,$$

where the middle group is the isomorphism classes of injective holomorphic $\mathbb{C}^*$ actions over $Q \setminus D$, δ sends such an isomorphism class to its "Chern class" and $\operatorname{Pic}(Q \setminus D)$ represents the deformations. As before, $H^2(Q;\mathbb{Z}) \cong \mathbb{Z} \oplus \text{Torsion}$.

In the discussion above, we have fixed $i:\mathbb{Z} \to \mathbb{C}$ and $\rho: Q \to \operatorname{Hol}(D)$. If we vary these choices, we don't get anything new in the smooth case because of smooth rigidity. That is, $\theta(\pi)$ is conjugate to $\theta'(\pi)$ in $\operatorname{Diff}_{\mathbb{C}}(\mathbb{C} \times D)$ where conjugation is taken in the whole group and not just in $\mathcal{C}(D,\mathbb{C})$ as for strict equivalence. However, in the holomorphic case, a change in $\rho: Q \to \operatorname{Hol}(D)$ induces a much larger deformation space than treated above. We can see this in our next example of reduction of the universal group where instead of considering complex structures and complex actions, we replace them by essentially equivalent Riemannian metric structures and metric preserving S^1-actions on N^3 $(M = N \times \mathbb{R}^1)$.

6.3. $\widetilde{\mathrm{PSL}}(2,\mathbb{R})$-*geometry*

6.3.1. When W is homeomorphic to $\mathbb{R}^2$ in a Seifert Construction and $\rho(Q)$ is a discrete subgroup of $\mathrm{TOP}(\mathbb{R}^2)$, acting properly on $\mathbb{R}^2$, then the group $\rho(Q)$ can be conjugated in $\mathrm{TOP}(\mathbb{R}^2)$ to a group which acts as isometries on $\mathbb{R}^2$ with the usual Euclidean metric (e.g., $\rho(Q)$ is crystallographic) or as isometries on $\mathbb{R}^2$ with the usual hyperbolic metric. In the former case, we say $\rho(Q)$ is isomorphic to a (Euclidean) crystallographic group, and in the latter, $\rho(Q)$ is isomorphic to a hyperbolic group. We write $\mathbb{R}^2$ with the usual Euclidean metric (resp. hyperbolic metric) as $\mathbb{E}^2$ (resp. $\mathbb{H}$). If $Q \setminus \mathbb{R}^2$ is compact, then any two embeddings of Q are conjugate in $\mathrm{TOP}(\mathbb{R}^2)$. However, if $\rho_1, \rho_2 : Q \to \mathrm{Isom}(\mathbb{H})$ have compact quotients, then ρ_1 is conjugated in $\mathrm{Isom}(\mathbb{H})$ to ρ_2 if and only if $\rho_2(Q)$ lies in the normalizer of $\rho_1(Q)$ in $\mathrm{Isom}(\mathbb{H})$ (the normalizer is a finite extension of $\rho_1(Q)$). Thus, if we reduce the universal group $\mathrm{TOP}_G(G \times \mathbb{R}^2)$ to $\mathcal{U}$, where at least $\mathrm{TOP}(\mathbb{R}^2)$ is replaced by $\mathrm{Isom}(\mathbb{E}^2)$ or $\mathrm{Isom}(\mathbb{H})$, we would expect to find a rich deformation theory for Seifert Constructions modelled on $G \times \mathbb{R}^2$. We will take $G = \mathbb{R}$, and $\overline{\mathcal{U}} = \mathrm{Isom}(\mathbb{H}) \subset \mathrm{TOP}(\mathbb{R}^2)$.

For each central extension $0 \to \mathbb{Z} \to \pi \to Q \to 0$ with Q cocompact hyperbolic, the group π can be embedded by $\theta: \pi \to \mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{H})$ so that it is topologically and/or smoothly rigid. If π is torsion free, $\pi \setminus (\mathbb{R} \times \mathbb{H})$ is a classical closed Seifert 3-manifold N^3 and with a unique S^1 action up to equivalence. The S^1-orbit space or base space is a 2-dimensional orbifold isomorphic to $Q \setminus \mathbb{H}$. The product $\mathbb{R} \times \mathbb{H}$ carries several geometries so that the Riemannian metric induced on $\mathbb{R} \times \mathbb{H} \to \mathbb{H}$ is the hyperbolic metric. We will examine first the Riemannian metric on $\widetilde{\mathrm{PSL}}(2,\mathbb{R})$ and then later an $\mathbb{R}$-invariant Lorentz metric on $\mathbb{R} \times \mathbb{H}$.

Let us denote $\widetilde{\mathrm{PSL}}(2,\mathbb{R})$ by $\widetilde{P}$. The space $\widetilde{P}$ is the universal covering group of $\mathrm{PSL}(2,\mathbb{R})$. Topologically $\widetilde{P}$ is homeomorphic to $\mathbb{R}\times\mathbb{H}$.

$$\begin{aligned} P &= \mathrm{PSL}(2,\mathbb{R}),\\ \widetilde{P} &= \widetilde{\mathrm{PSL}}(2,\mathbb{R})\\ &\approx \mathbb{R}\times\mathbb{H}. \end{aligned}$$

Therefore, if we use

$$\mathrm{TOP}_{\mathbb{R}}(\mathbb{R}\times\mathbb{H}) = \mathrm{M}(\mathbb{H},\mathbb{R}) \rtimes \big(\mathrm{GL}(1,\mathbb{R})\times \mathrm{TOP}(\mathbb{H})\big)$$

as our universal group, we will not be able to distinguish the geometries of $\mathbb{R}\times\mathbb{H}$, $\mathbb{R}\times\mathbb{R}^2$ and Nil from that of $\widetilde{P}$.

The Lie group $\mathrm{PSL}(2,\mathbb{R})$ can be viewed as the unit tangent bundle of the hyperbolic space $\mathbb{H}$, and it has a natural Riemannian metric. This metric pulls back to a Riemannian metric on $\widetilde{P}$. It turns out that this metric is right invariant. That is, all right translations by elements of the group are isometries. Furthermore, the isometry group is

$$\mathrm{Isom}(\widetilde{P}) = (\mathbb{R}\times_{\mathbb{Z}}\widetilde{P}) \rtimes \mathbb{Z}_2,$$

where $\mathbb{R}$ is a subgroup of $\widetilde{P}$ containing the center $\mathbb{Z}$, acting as left translations. These two actions commute with each other, and $\hat{\ell}(z) = \hat{r}(z^{-1})$ for $z\in\mathbb{Z}$, the center of $\widetilde{P}$. The finite group $\mathbb{Z}_2$ is generated by the reflection about the y-axis. In fact, any orientation-reversing isometry of period 2 will do. While it reverses the orientation of the base space $\mathbb{H}$, it also reverses the orientation of the fiber $\mathbb{R}$. Consequently, it preserves the orientation of $\widetilde{P}$.

We take the subgroup $\mathbb{R}$ described above as our G. Smoothly, $\widetilde{P} = \mathbb{R}\times\mathbb{H}$. Therefore, we have

$$G = \mathbb{R}, \qquad W = \mathbb{H}.$$

Since $\mathbb{R}\times_{\mathbb{Z}}\widetilde{P}$ commutes with the left translation $G=\mathbb{R}$, and the generator of $\mathbb{Z}_2$ is an inversion of $\mathbb{R}$, $\mathrm{Isom}(\widetilde{P}) = (\mathbb{R}\times_{\mathbb{Z}}\widetilde{P})\rtimes\mathbb{Z}_2$ lies inside $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R}\times\mathbb{H})$. To make the presentation clearer, we use only the connected component of $\mathrm{Isom}(\widetilde{P})$. So, let's take

$$\mathcal{U} = \mathrm{Isom}_0(\widetilde{P}) = \mathbb{R}\times_{\mathbb{Z}}\widetilde{P}$$

so that we have the commuting diagram:

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathcal{U}=\mathbb{R}\times_{\mathbb{Z}}\widetilde{P} & \longrightarrow & P & \longrightarrow & 1\\ & & \cap & & \cap & & \cap & & \\ 1 & \longrightarrow & \mathrm{M}(\mathbb{H},\mathbb{R}) & \longrightarrow & \mathrm{TOP}_{\mathbb{R}}(\mathbb{R}\times\mathbb{H}) & \longrightarrow & \mathrm{GL}(1,\mathbb{R})\times\mathrm{TOP}(\mathbb{H}) & \longrightarrow & 1 \end{array}$$

For this case, Q coincides with the Q examined in Section 6.2, and has a well known presentation

$$Q = \left\langle \bar{x}_1, \ldots, \bar{x}_g, \bar{y}_1, \ldots, \bar{y}_g, \overline{w}_1, \ldots, \overline{w}_p \mid \overline{w}_j^{\alpha_j} = 1, \prod_{j=1}^{p} \overline{w}_j \prod_{i=1}^{g} [\bar{x}_i, \bar{y}_i] = 1 \right\rangle$$

for $p \geqslant 0$, $g \geqslant 0$ and all $\alpha_j \geqslant 2$. It is also required that the Euler characteristic of Q, defined as

$$\chi(Q) = (2 - 2g) - \sum_{j=1}^{p} \left(1 - \frac{1}{\alpha_j}\right),$$

satisfies $\chi(Q) < 0$. It is our intention to characterize those $[\pi] \in H^2(Q; \mathbb{Z})$ which embed in $\mathcal{U}$ and to determine their deformation spaces.

6.3.2. For any subgroup Q of $\mathrm{Isom}_0(\mathbb{H})$, one can pullback (see Subsection 3.3.2) the above extension via $Q \hookrightarrow \mathrm{Isom}_0(\mathbb{H})$ to get $\widetilde{Q}$ so that the diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \mathbb{R} & \longrightarrow & \widetilde{Q} & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{=} & & \downarrow & & \downarrow{\cap} & & \\ 1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathrm{Isom}_0(\widetilde{P}) & \longrightarrow & \mathrm{Isom}_0(\mathbb{H}) & \longrightarrow & 1 \end{array}$$

commutes. Thus, $\mathbb{R}$ becomes a (trivial) Q-module.

LEMMA 6.3.3. *Let Q be a cocompact discrete subgroup of* $\mathrm{Isom}_0(\mathbb{H})$. *Then*

(1) $H^2(Q; \mathbb{R}) = \mathbb{R}$.

(2) *The class $[\widetilde{Q}] \in H^2(Q; \mathbb{R})$ is non-zero. (That is, the exact sequence does not split.)*

PROOF. (1) Let $1 \to \mathbb{Z} \to \widetilde{P} \to P \to 1$ be the universal covering projection, and let $1 \to \mathbb{Z} \to \widehat{Q} \to Q \to 1$ be the pullback of this exact sequence via $Q \hookrightarrow P$. Then $\mathbb{Z} \subset \widehat{Q}$ sits in $\mathbb{R} \times_{\mathbb{Z}} \widetilde{P} = \mathrm{Isom}_0 \widetilde{P}$ as the center. Denote the inclusion of $\mathbb{Z} \hookrightarrow \mathbb{R} \subset \mathrm{Isom}_0(\widetilde{P})$ by i. Then

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \mathbb{Z} & \longrightarrow & \widehat{Q} & \longrightarrow & Q & \longrightarrow & 1 \\ & & \downarrow{i} & & \downarrow & & \downarrow{=} & & \\ 1 & \longrightarrow & \mathbb{R} & \longrightarrow & \widetilde{Q} & \longrightarrow & Q & \longrightarrow & 1 \end{array} \qquad \begin{array}{c} [\widehat{Q}] \in H^2(Q; \mathbb{Z}) \\ \downarrow{i_*} \\ [\widetilde{Q}] \in H^2(Q; \mathbb{R}) \end{array}$$

is commutative so that $i_*[\widehat{Q}] = [\widetilde{Q}]$. By Selberg's lemma, Q contains a torsion free normal subgroup Q_0 of finite index. Then $H^2(Q_0; \mathbb{R}) = H^2(Q_0 \setminus \mathbb{H}; \mathbb{R}) \cong \mathbb{R}$. Then, by transfer, $H^2(Q; \mathbb{R}) = \mathbb{R}$, since Q/Q_0 acts on $Q_0 \setminus \mathbb{H}$ preserving orientation. Since $H^2(Q; \mathbb{Z})$ is finitely generated, it follows that $H^2(Q; \mathbb{Z}) = \mathbb{Z} \oplus \mathrm{Torsion}$ by the Universal Coefficient Theorem, and the fact that $i_* : H^2(Q; \mathbb{Z}) \to H^2(Q; \mathbb{R})$ is given by $\bigotimes \mathbb{R}$. Thus the elements of infinite order inject and those of finite order are in the kernel.

(2) We can assume, without loss of generality, $Q \subset P$ is torsion free. In [55], it is shown that $[\widehat{Q}] \in H^2(Q; \mathbb{Z})$ is non-zero. In fact, $\widehat{Q}$ is the fundamental group of the unit tangent

bundle $\widehat{Q}\setminus\widetilde{P}$ of the surface $Q\setminus\mathbb{H}$. In this case, the Euler characteristic of Q, $2-2g\neq 0$, is the characteristic class of the principal S^1-bundle $\widehat{Q}\setminus\widetilde{P}\to Q\setminus\mathbb{H}$, and is also equal to the negative of the cohomology class $[\widehat{Q}]\in H^2(Q;\mathbb{Z})\cong\mathbb{Z}$ of the extension $\widehat{Q}$. Therefore, $[\widehat{Q}]$ is non-zero in $H^2(Q;\mathbb{R})=H^2(Q;\mathbb{Z})\otimes\mathbb{R}$. We also point out, in this case, $[\widehat{Q}]=e(\widehat{Q})$, see Subsection 6.3.4 for definition of $e(\widehat{Q})$. □

6.3.4. (Euler number) Let $i:\mathbb{Z}\hookrightarrow\mathbb{R}$ be the standard inclusion. For each central extension $0\to\mathbb{Z}\to\pi\to Q\to 1$, there is associated a rational invariant called the *Euler number* of π, and is denoted by $e(\pi)$. It can be defined in terms of a presentation of π. Let

$$\pi=\Big\langle \tilde{x}_1,\dots,\tilde{x}_g,\tilde{y}_1,\dots,\tilde{y}_g,\widetilde{w}_1,\dots,\widetilde{w}_p,\tilde{z}\mid \tilde{z}\text{ central},\\ \widetilde{w}_j^{\alpha_j}=\tilde{z}^{-\beta_j},\ \prod_{j=1}^{p}\widetilde{w}_j\prod_{i=1}^{g}[\tilde{x}_i,\tilde{y}_i]=\tilde{z}^b\Big\rangle.$$

Then $e(\pi)=-(b+\Sigma\frac{\beta_j}{\alpha_j})$ and $|e(\pi)|$ is an invariant of the isomorphism class of π. In [31, Theorem 4.5], it is shown that under the homomorphism

$$i_*:H^2(Q;\mathbb{Z})\longrightarrow H^2(Q;\mathbb{R})\cong\mathbb{R}$$

induced by i, we have $i_*[\pi]=L\cdot e(\pi)$, where $L=\mathrm{lcm}[\alpha_1,\dots,\alpha_p]$. Thus $[\pi]$ has infinite order in $H^2(Q;\mathbb{Z})$ if and only if $e(\pi)\neq 0$.

We now characterize the cocompact orbifold groups modelled on $\widetilde{P}$.

THEOREM 6.3.5. *An abstract group π can be embedded into* $\mathrm{Isom}_0(\widetilde{P})$ *as a cocompact discrete subgroup if and only if π is a central extension of $\mathbb{Z}$ by a discrete cocompact orientation-preserving hyperbolic group Q (so that $1\to\mathbb{Z}\to\pi\to Q\to 1$ is exact) and $[\pi]\in H^2(Q;\mathbb{Z})$ has infinite order. Further, if this is the case, the subgroup $\mathbb{Z}$ is the center of π, and in any discrete embedding, the image of $\mathbb{Z}$ is $\pi\cap\mathbb{R}$, where $\mathbb{R}\subset\mathbb{R}\times_{\mathbb{Z}}\widetilde{P}=$* $\mathrm{Isom}_0(\widetilde{P})$.

PROOF. Suppose π is a cocompact discrete subgroup of $\mathrm{Isom}_0(\widetilde{P})$. The subgroup $\mathbb{R}$ of $\mathrm{Isom}_0(\widetilde{P})$ is the radical (maximal connected normal solvable subgroup) of $\mathrm{Isom}_0(\widetilde{P})$ and the quotient $\mathrm{Isom}_0(\mathbb{H})=\mathrm{PSL}(2,\mathbb{R})$ has no compact factor. A theorem of Wang (see [51, 8.27]) says that the image Q of π in $\mathrm{Isom}_0(\mathbb{H})$ is a lattice so that Q is a discrete cocompact orientation-preserving hyperbolic group. It remains to show that $\pi\cap\mathbb{R}$ is non-trivial. Suppose not. Then π is isomorphic to Q and hence, it has $\mathbb{R}$-cohomological dimension 2. However, since π is cocompact in $\mathrm{Isom}_0(\widetilde{P})$, its $\mathbb{R}$-cohomological dimension is 3. This contradiction shows that $\pi\cap\mathbb{R}=\mathbb{Z}$. Thus π is of the form $1\to\mathbb{Z}\to\pi\to Q\to 1$. Clearly, $\mathbb{Z}$ is the center of π since Q is centerless.

We shall now sketch why $[\pi]$ must be of infinite order in $H^2(Q;\mathbb{Z})$. By Selberg's lemma, there exists a normal subgroup Q_0 of Q which is torsion free of finite index. Let $1\to\mathbb{Z}\to\pi_0\to Q_0\to 1$ be the pullback of the above exact sequence via $Q_0\hookrightarrow Q$. But if

$[\pi]$ has finite order, one can take Q_0 so that $[\pi_0]$ has order 0 so that $\pi_0 = \mathbb{Z} \times Q_0$. We claim that this group does not embed discretely into $\mathrm{Isom}_0(\widetilde{P})$. Choose a standard presentation for Q_0:

$$\left\langle a_1, b_1, \ldots, a_g, b_g \mid \prod_{i=1}^{g} [a_i, b_i] = 1 \right\rangle.$$

Since $Q_0 \subset \mathrm{PSL}(2, \mathbb{R})$, we may think of the a_i, b_i as elements of $\mathrm{PSL}(2, \mathbb{R})$. In $\mathrm{Isom}_0(\widetilde{P})$, these elements lift to $\{(a_i', t_{a_i})\}$, $\{(b_i', t_{b_i})\}$, were $a_i', b_i' \in \widetilde{P}$; $t_{a_i}, t_{b_i} \in \mathbb{R}$. These are unique up to the center of $\widetilde{P}$. Since $\prod_{i=1}^{g} [(a_i', t_{a_i}), (b_i', t_{b_i})] = (t^{2g-2}, 0)$ by [55], it is non-zero. Since $\pi_0 = \mathbb{Z} \times Q_0 \subset \mathrm{Isom}_0(\widetilde{P})$, this relation $2g - 2$ would have to be 0. This gives a contradiction and so $[\pi]$ must have infinite order.

Conversely, suppose $1 \to \mathbb{Z} \to \pi \to Q \to 1$ is exact, where Q is a cocompact discrete subgroup of $\mathrm{Isom}_0(\mathbb{H}) = \mathrm{PSL}(2, \mathbb{R})$; and $[\pi] \in H^2(Q; \mathbb{Z})$ has infinite order. Then with the natural inclusion $i : \mathbb{Z} \hookrightarrow \mathbb{R}$ and the induced homomorphism $i_* : H^2(Q; \mathbb{Z}) \to H^2(Q; \mathbb{R})$, $i_*[\pi]$ is the pushout $[\mathbb{R}\pi] \in H^2(Q; \mathbb{R})$ (see Section 3.3.1), and is non-zero. By Lemma 6.3.3, $[\widetilde{Q}] \in H^2(Q; \mathbb{R})$ is also non-zero. Therefore, there exists $\varepsilon \in \mathbb{R}$ so that $(\varepsilon \circ i)_*[\pi] = \varepsilon_*[\mathbb{R}\pi] = [\widetilde{Q}]$. This implies that there exists a homomorphism of π into $\widetilde{Q}$ with the diagram

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{Z} & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle i} & & \downarrow & & \downarrow{\scriptstyle =} & & \\
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathbb{R}\pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \varepsilon} & & \downarrow & & \downarrow{\scriptstyle =} & & \\
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \widetilde{Q} & \longrightarrow & Q & \longrightarrow & 1
\end{array}$$

commutative and with injective vertical maps. Since $\mathbb{Z}$ and Q acts properly discontinuously with compact quotient on $\mathbb{R}$ and $\mathbb{H}$ respectively, π is cocompact and discrete. This completes the proof. □

COROLLARY 6.3.6. *Let $\rho : Q \to \mathrm{PSL}(2, \mathbb{R})$ be a discrete cocompact subgroup. For an extension $1 \to \mathbb{Z} \to \pi \to Q \to 1$, there exists an injective homomorphism $\theta : \pi \to \mathrm{Isom}(\widetilde{P})$ so that the diagram*

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{Z} & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \varepsilon} & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle \rho} & & \\
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathrm{Isom}_0(\widetilde{P}) & \longrightarrow & \mathrm{PSL}(2, \mathbb{R}) & \longrightarrow & 1
\end{array} \tag{6.3}$$

commutes if and only if $[\pi] \in H^2(Q; \mathbb{Z})$ has infinite order.

COROLLARY 6.3.7. (Structure) *Let M be a closed orbifold modelled on* $(\mathrm{Isom}_0(\widetilde{P}), \widetilde{P})$-*geometry. Then M is an orientable closed Seifert orbifold over a hyperbolic base with* $e(M) \neq 0$.

(Realization) *Let M be a compact orientable Seifert orbifold over a hyperbolic orbifold. Then M admits an* $(\mathrm{Isom}_0(\widetilde{P}), \widetilde{P})$-*geometry if and only if* $e(M) \neq 0$.

6.3.8. Consider the commuting diagram (6.3). With fixed ε and ρ, how many θ's are there to make the diagram commutative? Such maps are classified by $H^1(Q; \mathbb{R})$, see Theorem 3.7.2. Since the action of Q on $\mathbb{R}$ is trivial,

$$H^1(Q; \mathbb{R}) = \mathbb{R}^{2g},$$

where $2g$ is the first Betti number of the group Q. Compare this with

$$H^i\big(Q; \mathrm{M}(\mathbb{H}, \mathbb{R})\big) = 0 \quad (i \geqslant 1)$$

so that, for any extension $1 \to \mathbb{Z} \to \pi \to Q \to 1$, a homomorphism $\pi \to \mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{H})$ exists and is unique, up to conjugation by elements of $\mathrm{M}(\mathbb{H}, \mathbb{R})$.

6.4. *Lorentz structures and* $\widetilde{\mathrm{PSL}(2, \mathbb{R})}$-*geometry*

The spaces admitting $\widetilde{\mathrm{PSL}_2\mathbb{R}}$-geometry have another interesting geometric structure. Here is a more explicit description of our problem. Let us denote $\widetilde{\mathrm{PSL}}(2, \mathbb{R})$ by P_∞. The space P_∞ is the universal covering group of $P_1 = \mathrm{PSL}(2, \mathbb{R})$. Topologically P_∞ is homeomorphic to $\mathbb{R} \times \mathbb{H}$.

$$P_1 = \mathrm{PSL}(2, \mathbb{R}),$$
$$P_\infty = \widetilde{\mathrm{PSL}}(2, \mathbb{R}).$$

Consider the *indefinite* metric of signature $++--$ on $\mathbb{R}^4$. The unit sphere of this space is

$$S^{1,2} = \big\{(x, y) \mid x, y \in \mathbb{R}^2,\ |x|^2 - |y|^2 = 1\big\} \approx \mathrm{O}(2,2)/\mathrm{O}(1,2).$$

The linear map of $\mathbb{R}^4$ defined by the matrix

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}$$

transforms $S^{1,2}$ to

$$P_2 = \big\{(x, y, z, u) \in \mathbb{R}^4 \mid xu - yz = 1\big\} = \mathrm{SL}_2\,\mathbb{R}.$$

Thus P_2 has a complete *Lorentz metric* of signature $+, -, -$ and *constant sectional curvature* $= 1$.

A space is called a *Lorentz orbifold* if it is the quotient of P_∞ by a discrete group of Lorentz isometries acting *properly discontinuously*. One can show that such a group contains normal subgroups of finite index which act freely. A Lorentz orbifold for which the discrete group acts freely is called a *Lorentz space-form*. The Lorentz structure on a space-form is non-singular and it has P_∞ as its (metric) universal covering. Then $1 \to \mathbb{Z} \to P_\infty \to P_1 \to 1$ is actually a central extension with $\mathbb{Z}$ being the entire center of P_∞. It turns out that the identity component of $\mathrm{Isom}(P_\infty)$ is $(P_\infty \times P_\infty)/\mathbb{Z}$ where $\mathbb{Z}$ is the diagonal central subgroup corresponding to the center of each of the P_∞-factors. The action $P_\infty \times P_\infty$, as isometries, on P_∞ is given by

$$(\alpha, \beta) \cdot x = \alpha \cdot x \cdot \beta^{-1}.$$

Moreover, $\mathrm{Isom}_0(P_\infty) = P_\infty \times_{\mathbb{Z}} P_\infty$ has index 4 in $\mathrm{Isom}(P_\infty)$. These Lorentz space-forms are analogous to the complete spherical space-forms in the Riemannian case.

Let us describe some obvious ones which turn out to be *homogeneous* in the sense that $\mathrm{Isom}(M)$ acts *transitively* on M. Take $\pi \subset P_\infty \times e \subset P_\infty \times_{\mathbb{Z}} P_\infty$ as a discrete subgroup. Then surely the centralizer of π in $\mathrm{Isom}(P_\infty)$, $C_{\mathrm{Isom}(P_\infty)}(\pi)$, contains $e \times P_\infty$ in $(P_\infty \times_{\mathbb{Z}} P_\infty)$. In fact, it is exactly $e \times P_\infty$ (unless $\pi \approx \mathbb{Z}$ and sits in the center of $P_\infty \times e$ in $\mathrm{Isom}(P_\infty)$). Such groups π are classified in [55] and are certain Seifert manifolds over a hyperbolic base, for there is an obvious S^1 action on $\pi \setminus P_\infty$ induced from $e \times P_\infty \subset \mathrm{Isom}_0(P_\infty)$. A surprising fact is that *all homogeneous Lorentz orbifolds are* actually *homogeneous Lorentz space forms and coincide with those just described above*, [30, §10].

If $\pi \subset \mathrm{Isom}(P_\infty)$ so that $M = \pi \setminus P_\infty$ is compact then it is shown in [30, §7] that M *is homeomorphic to an orientable Seifert orbifold over a hyperbolic base.* However, the connections between the Seifert structure and the Lorentz structure is unclear. This is due to the fact that $\mathrm{Isom}(P_\infty)$ does not act properly on P_∞. By selecting a maximal subgroup of $\mathrm{Isom}(P_\infty)$ which acts properly on P_∞, these two disparate structures can be related.

The subgroup

$$J(P_\infty) = (P_\infty \times_{\mathbb{Z}} \mathbb{R}) \rtimes \mathbb{Z}_2 \subset \mathrm{Isom}(P_\infty),$$

where $P_\infty \times_{\mathbb{Z}} \mathbb{R} = (P_\infty \times \mathbb{R})/\mathbb{Z}$, where $\mathbb{Z}$ is the central diagonal subgroup of $P_\infty \times \mathbb{R}$, and $\mathbb{Z}_2$ reverses the orientation of time ($= \mathbb{R}$) and space ($= \mathbb{H} = P_\infty/\mathbb{R}$) at the same time is the same Lie group as the $\mathrm{Isom}(P_\infty)$ in the Riemannian case. A *Lorentz orbifold* (resp. *space-form*) $M = \pi \setminus P_\infty$, where $\pi \subset J(P_\infty) \subset \mathrm{Isom}(P_\infty)$ is called *standard*. Therefore the standard Lorentz orbifolds and space-forms coincide with the orbifolds and space-forms of the $\widetilde{\mathrm{PSL}(2,\mathbb{R})}$-geometry. A Lorentz space-form is homogeneous if the full group of Lorentz isometries acts transitively. It is known that homogeneous Lorentz space-forms M admit nonstandard complete Lorentz structures if $H^1(M; \mathbb{R}) \neq 0$. See [18].

The *Seifert fibering*, in the theorem below, on M *descends* from the $\mathbb{R}$-action by the second $\mathbb{R}$-factor in $J(P_\infty)$ on P_∞. The following are due to Kulkarni and Raymond and are stated here, for simplicity, in the *closed cases*.

THEOREM 6.4.1 ([30, (8.5)]). (Structure) *A compact standard Lorentz space-form (resp. Lorentz orbifold) is an orientable Seifert manifold M (resp. Seifert orbifold) over a hyperbolic base B with $e(M) \neq 0$.*

(Realization). *Let M be a compact orientable Seifert manifold (resp. orientable Seifert orbifold) over a hyperbolic base with $e(M) \neq 0$. Then, M admits a structure of a standard Lorentz space-form (resp. Lorentz orbifold).*

We remark that the orbifold part breaks into two separate cases. If all fibers are $\approx S^1$, then *all* closed orientable Seifert *manifolds* with $e(M) \neq 0$ appear as Lorentz *orbifolds* (and conversely). This is similar to having the topological sphere appear as a 2-dimensional hyperbolic orbifold. If some fibers are arcs then the Lorentz orbifolds are homeomorphic to connected sums of lens spaces (including S^3 and $S^2 \times S^1$). We should also mention that the statements in Theorem 6.4.1 are for the full $J(P_\infty)$ and not the connected component of the identity, and correspond to the cocompact discrete subgroups of $\mathrm{Isom}(\widetilde{P})$ in the Riemannian case instead of $\mathrm{Isom}_0(\widetilde{P})$ as described in the preceding section. See [30, §8, 9] for details and treatment of cases other than the compact ones.

6.5. *Deformation spaces for* $\widetilde{\mathrm{PSL}}(2,\mathbb{R})$*-geometry*

DEFINITION 6.5.1. Let $\mathcal{R}(\pi;\mathcal{U})$ be the space of all injective homomorphisms $\theta:\pi \to \mathcal{U}$ such that $\theta(\pi)$ is cocompact acting properly on P_∞ and, is discrete in $\mathcal{U}$. We topologize $\mathcal{R}(\pi;\mathcal{U})$ as a subset of $\mathcal{U}^\pi$. In general, if $\mathcal{U}$ is a Lie group, then $\mathcal{R}(\pi;\mathcal{U})$ will be a real analytic space.

The space $\mathcal{R}(\pi;\mathcal{U})$ is called *the space of discrete representations of π into $\mathcal{U}$* or *the Weil space of* $(\pi;\mathcal{U})$. When there is no confusion likely, we denote $\mathcal{R}(\pi;\mathcal{U})$ simply by $\mathcal{R}(\pi)$.

Recall that μ denotes conjugation. The inner-automorphisms group $\mathrm{Inn}(\mathcal{U})$ acts on $\mathcal{R}(\pi)$ from the left by

$$\mu(u)\cdot\theta = \mu(u)\circ\theta$$

for $u \in \mathcal{U}$ and $\theta \in \mathcal{R}(\pi)$. Denote the orbit space of this action by

$$\mathcal{T}(\pi) = \mathrm{Inn}(\mathcal{U}) \setminus \mathcal{R}(\pi).$$

It is called the *Teichmüller space* of π (or of M).

$\mathrm{Aut}(\pi)$ acts on $\mathcal{R}(\pi)$ from the right by

$$\theta\cdot f = \theta\circ f$$

for $\theta \in \mathcal{R}(\pi)$ and $f \in \mathrm{Aut}(\pi)$. Denote the orbit space of this action by

$$\mathcal{S}(\pi) = \mathcal{R}(\pi)/\mathrm{Aut}(\pi).$$

$\mathcal{S}(\pi)$ is the *space of discrete subgroups of $\mathcal{U}$ each isomorphic to π*, or the *Chabauty space*.

Since the two actions of $\mathrm{Inn}(\mathcal{U})$ and $\mathrm{Aut}(\pi)$ commute with each other, $\mathrm{Aut}(\pi)$ acts on $\mathcal{T}(\pi)$, and $\mathrm{Inn}(\mathcal{U})$ acts on $\mathcal{S}(\pi)$. $\mathrm{Aut}(\pi)$ has an obvious kernel $\mathrm{Inn}(\pi)$. Consequently, we get an action of $\mathrm{Out}(\pi)$ on $\mathcal{T}(\pi)$. We denote the orbit space by

$$\mathcal{M}(\pi) = \mathcal{T}(\pi)/\mathrm{Out}(\pi).$$

It is called the *moduli* (or *Riemann*) *space* of π. It is also obtained as the orbit space

$$\mathcal{M}(\pi) = \mathrm{Inn}(\mathcal{U}) \setminus \mathcal{S}(\pi).$$

Summarizing in the form of a commutative diagram of orbit mappings, we have

$$\begin{array}{ccc} (\mathrm{Inn}(\mathcal{U}), \mathcal{R}(\pi), \mathrm{Aut}(\pi)) & \xrightarrow{\mathrm{Inn}(\mathcal{U})\setminus} & (\mathcal{T}(\pi), \mathrm{Out}(\pi)) \\ \Big\downarrow{\scriptstyle /\mathrm{Aut}(\pi)} & & \Big\downarrow{\scriptstyle /\mathrm{Out}(\pi)} \\ (\mathrm{Inn}(\mathcal{U}), \mathcal{S}(\pi)) & \xrightarrow{\mathrm{Inn}(\mathcal{U})\setminus} & \mathcal{M}(\pi) \end{array} \tag{6.4}$$

We shall now describe the deformation spaces for closed M which have a geometric structure modelled on $(\mathcal{U}, \widetilde{P}) = (\mathrm{Isom}_0(\widetilde{P}), \widetilde{P})$ (or equivalently the *standard* Lorentz structures). We shall see that these deformation spaces all have Seifert fiberings over well-studied deformation spaces of discrete cocompact orientation-preserving hyperbolic groups.

Let π be a cocompact discrete subgroup of $\mathcal{U} = \mathrm{Isom}_0(P_\infty)$. We have the central extension

$$1 \to \mathbb{Z} \to \pi \to Q \to 1,$$

$[\pi] \in H^2(Q;\mathbb{Z})$ with $Q \subset P$, and having infinite order in $H^2(Q;\mathbb{Z})$.

Since our group $\mathcal{U}$ embeds into $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{H})$ we have *topological rigidity* in the *strong sense* that if θ_1 and θ_2 are two embeddings of cocompact π into $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{H})$, then they are conjugate in $\mathrm{TOP}_{\mathbb{R}}(\mathbb{R} \times \mathbb{H})$. In almost all cases they will not be conjugate in $\mathcal{U}$. The elements of $\mathcal{M}(\pi)$ then represent the different $\mathcal{U}$-structures on M, and we may expect large deformation spaces.

In order to understand $\mathcal{R}(\pi)$, we need to study $\mathrm{Aut}(\pi)$. We would like to describe $\mathrm{Aut}(\pi)$ in terms of $\mathrm{Aut}(Q)$. Since $\mathbb{Z}$ is characteristic, any automorphism of π induces an automorphism of Q.

DEFINITION 6.5.2. Let $\mathrm{Aut}(Q(\pi))$ be the image of $\mathrm{Aut}(\pi) \to \mathrm{Aut}(Q)$. That is,

$$\mathrm{Aut}\big(Q(\pi)\big) = \big\{\bar{\theta} \in \mathrm{Aut}(Q)\colon \exists\theta \in \mathrm{Aut}(\pi) \text{ inducing } \bar{\theta}\big\},$$

the group of automorphisms of Q which can be lifted to an automorphism of π.

Because $\mathbb{Z} \subset \mathbb{R}$ has the unique isomorphism extension property, one can form a pushout (see Section 3.3.1) to get $\mathbb{R}\pi$ fitting the commuting diagram:

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{Z} & \longrightarrow & \pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathbb{R}\pi & \longrightarrow & Q & \longrightarrow & 1
\end{array}$$

LEMMA 6.5.3. *There is a commutative diagram with exact rows and injective vertical maps*:

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathrm{Hom}(Q,\mathbb{Z}) & \longrightarrow & \mathrm{Aut}(\pi) & \longrightarrow & \mathrm{Aut}(Q(\pi)) & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \mathrm{Hom}(Q,\mathbb{R}) & \longrightarrow & \mathrm{Aut}(\mathbb{R}\pi) & \longrightarrow & \mathrm{Aut}(Q) & \longrightarrow & 1
\end{array} \tag{6.5}$$

PROOF. The crucial fact for the proof is $H^2(Q;\mathbb{R}) = \mathbb{R}$. Since $[\pi] \in H^2(Q;\mathbb{Z})$ has infinite order, $[\mathbb{R}\pi] \in H^2(Q;\mathbb{R})$ is non-zero. Since $\mathbb{R}$ is characteristic in $\mathbb{R}\pi$, any $f \in \mathrm{Aut}(\mathbb{R}\pi)$ induces an automorphism $\bar{f} \in \mathrm{Aut}(Q)$. Suppose $\bar{f} = \mathrm{id}$. Let $\hat{f}:\mathbb{R} \to \mathbb{R}$ be the restriction of f. Then $\hat{f}_*[\mathbb{R}\pi] = \bar{f}^*[\mathbb{R}\pi] = [\mathbb{R}\pi]$ since $\bar{f} = \mathrm{id}$. Since $[\mathbb{R}\pi]$ is non-zero, $\hat{f} = \mathrm{id}$. Therefore, f is of the form

$$f(\alpha) = \lambda(\alpha) \cdot \alpha$$

for some map $\lambda : \pi \to \mathbb{R}$. One easily sees that λ factors through Q and it satisfies the cocycle condition

$$\lambda(\bar{\alpha}\bar{\beta}) = \lambda(\bar{\alpha}) + \bar{\alpha}\lambda(\bar{\beta})$$

for all $\bar{\alpha}$, $\bar{\beta} \in Q$ so that $\lambda \in Z^1(Q,\mathbb{R})$. However, since $\mathbb{R}$ is central in $\mathbb{R}\pi$, $\mathbb{R}$ is a trivial Q-module so that $Z^1(Q,\mathbb{R}) = \mathrm{Hom}(Q,\mathbb{R})$. Conversely, any such a $\lambda \in \mathrm{Hom}(Q,\mathbb{R})$ yields an automorphism $f \in \mathrm{Aut}(\mathbb{R}\pi)$. Moreover, $\mathrm{Hom}(Q,\mathbb{R}) \cap \mathrm{Aut}(\pi) = \mathrm{Hom}(Q,\mathbb{Z})$.

Let $\bar{g} \in \mathrm{Aut}(Q)$. We would like to find $g \in \mathrm{Aut}(\mathbb{R}\pi)$ which induces $\bar{g}$ on Q. The automorphism $\bar{g}$ induces an automorphism $\bar{g}^* : H^2(Q;\mathbb{R}) \to H^2(Q;\mathbb{R})$. Since $H^2(Q;\mathbb{R}) = \mathbb{R}$ and $[\mathbb{R}\pi] \neq 0$, there is a real number ε for which $\bar{g}^*[\mathbb{R}\pi] = \varepsilon[\mathbb{R}\pi]$. This number ε is non-zero and can be viewed as an automorphism of $\mathbb{R}\pi$ such that

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathbb{R}\pi & \longrightarrow & Q & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle \varepsilon} & & \downarrow & & \downarrow{\scriptstyle \bar{g}} & & \\
1 & \longrightarrow & \mathbb{R} & \longrightarrow & \mathbb{R}\pi & \longrightarrow & Q & \longrightarrow & 1
\end{array}$$

is commutative. Thus we have shown that $\mathrm{Aut}(\mathbb{R}\pi) \to \mathrm{Aut}(Q)$ is surjective. □

The deformation spaces of π will be studied via those of Q. To this end, it is necessary to define the following:

DEFINITION 6.5.4. Let $\mathcal{R}(Q;\overline{\mathcal{U}})$ be the space of all injective homomorphisms $\bar{\theta}(Q)$ such that $\bar{\theta}(Q)$ is cocompact and discrete in $\overline{\mathcal{U}}$.

There is a left action of $\mathrm{Inn}(\overline{\mathcal{U}})$ on $\mathcal{R}(Q;\overline{\mathcal{U}})$, and also a right action of $\mathrm{Aut}(Q(\pi))$ on $\mathcal{R}(Q;\overline{\mathcal{U}})$. We then define

$$\mathcal{T}(Q;\overline{\mathcal{U}}) = \mathrm{Inn}(\overline{\mathcal{U}}) \setminus \mathcal{R}(Q;\overline{\mathcal{U}}),$$
$$\mathcal{S}\big(Q(\pi);\overline{\mathcal{U}}\big) = \mathcal{R}(Q;\overline{\mathcal{U}})/\mathrm{Aut}\big(Q(\pi)\big),$$
$$\mathcal{M}\big(Q(\pi);\overline{\mathcal{U}}\big) = \mathcal{T}(Q;\overline{\mathcal{U}})/\mathrm{Out}\big(Q(\pi)\big).$$

These are the *Teichmüller space*, the *restricted Chabauty space*, and the *restricted moduli space* of Q, respectively.

It is known that $\mathrm{Aut}(Q(\pi))$ is a subgroup of $\mathrm{Aut}(Q)$ of finite index, see [29, Appendix]. This implies that $\mathcal{S}(Q(\pi)) = \mathcal{R}(Q)/\mathrm{Aut}(Q(\pi))$ is a finite regular covering of $\mathcal{S}(Q)$. Also $\mathcal{M}(Q(\pi)) = \mathcal{T}(Q)/\mathrm{Out}(Q(\pi))$, where $\mathrm{Out}(Q(\pi)) = \mathrm{Aut}(Q(\pi))/\mathrm{Inn}(Q)$.

LEMMA 6.5.5. *The action of* $\mathrm{Aut}(\pi)$ *on* $\mathcal{R}(\pi;\mathcal{U})$ *extends to an action of* $\mathrm{Aut}(\mathbb{R}\pi)$ *on* $\mathcal{R}(\pi;\mathcal{U})$. *The subgroup* $Z^1(Q;\mathbb{R}) = H^1(Q;\mathbb{R}) = \mathrm{Hom}(Q,\mathbb{R})$ *acts on* $\mathcal{R}(\pi;\mathcal{U})$ *freely and properly. Moreover,* $\mathcal{R}(\pi;\mathcal{U})/\mathrm{Hom}(Q,\mathbb{R}) = \mathcal{R}(Q;\overline{\mathcal{U}})$.

PROOF. Clearly, an element $\theta \in \mathcal{R}(\pi;\mathcal{U})$ determines a homomorphism $\tilde{\theta}:\mathbb{R}\pi \to \mathcal{U}$ uniquely. The action of $\mathrm{Aut}(\mathbb{R}\pi)$ on $\mathcal{R}(\pi;\mathcal{U})$ is defined as follows: for $\theta \in \mathcal{R}(\pi;\mathcal{U})$ and $f \in \mathrm{Aut}(\mathbb{R}\pi)$,

$$\theta \cdot f = \tilde{\theta} \circ f|_{\pi}.$$

The image of π under $\tilde{\theta} \circ f$ is a discrete cocompact subgroup of $\mathcal{U}$ so that $\tilde{\theta} \circ f|_{\pi} \in \mathcal{R}(\pi;\mathcal{U})$.

Suppose $\theta, \theta' \in \mathcal{R}(\pi;\mathcal{U})$ induce the same representation $\bar{\theta} = \bar{\theta}' \in \mathcal{R}(Q;\overline{\mathcal{U}})$. Let $\tilde{\theta}, \tilde{\theta}':\mathbb{R}\pi \to \mathcal{U}$ be the homomorphisms induced from θ, θ' as described above. Since $\bar{\theta} = \bar{\theta}'$, the embeddings $\tilde{\theta}$ and $\tilde{\theta}'$ are related by $\tilde{\theta}'(\alpha) = \lambda(\alpha)\tilde{\theta}(\alpha)$ for some map $\lambda:\mathbb{R}\pi \to \mathbb{R} \subset \mathcal{U}$. Since $\tilde{\theta}$ and $\tilde{\theta}'$ must be equal on $\mathbb{R}$, the map λ factors through Q, and hence

$$\lambda : Q \longrightarrow \mathbb{R}.$$

The map λ satisfies the cocycle condition so that $\lambda \in \mathrm{Hom}(Q,\mathbb{R})$. Conversely, let $f \in \mathrm{Aut}(\mathbb{R}\pi)$ inducing the identity on Q. Then by reversing the order of arguments given above, one sees that θ and $\theta \circ f$ represent the same element of $\mathcal{R}(Q;\overline{\mathcal{U}})$.

Clearly, unless f is the identity, θ and $\theta \circ f$ will be different, which shows that the action of $\mathrm{Hom}(Q,\mathbb{R})$ on $\mathcal{R}(\pi;\mathcal{U})$ is free and proper since the orbit space is Hausdorff. □

THEOREM 6.5.6 ([29, Theorem 2.5]). *Let* π *be a compact orbifold group with* $(\mathrm{Isom}_0(\widetilde{P}), \widetilde{P})$*-geometry. Let* g *be the genus of the base orbifold. Then,*

$$\mathcal{R}(\pi) = \mathcal{R}(Q) \times \mathbb{R}^{2g} \ \textit{trivial principal } \mathbb{R}^{2g}\textit{-bundle over } \mathcal{R}(Q),$$

$\mathcal{T}(\pi) = \mathcal{T}(Q) \times \mathbb{R}^{2g}$ *trivial principal* $\mathbb{R}^{2g}$*-bundle over* $\mathcal{T}(Q)$,
$\mathcal{S}(\pi)$ T^{2g}*-bundle over* $\mathcal{S}(Q(\pi))$,
$\mathcal{M}(\pi)$ *Seifert fiber space over* $\mathcal{M}(Q(\pi))$ *with typical fiber* T^{2g}.

Furthermore, $\mathcal{S}(Q(\pi))$ *is a finite sheeted covering of* $\mathcal{S}(Q)$*, and* $\mathcal{M}(Q(\pi))$ *is a finite sheeted branched covering of* $\mathcal{M}(Q)$.

PROOF. Let $\mathbb{Z}$ be the center of π. Then $Q = \pi/\mathbb{Z}$ is the base orbifold group. Hence in Lemma 6.5.3, $\mathrm{Hom}(Q, \mathbb{R}) = H^1(Q; \mathbb{R}) = \mathbb{R}^{2g}$, and

$$1 \longrightarrow \mathbb{R}^{2g} \longrightarrow \mathrm{Aut}(\mathbb{R}\pi) \longrightarrow \mathrm{Aut}(Q) \longrightarrow 1$$

is exact. By Lemma 6.5.5, the group $\mathbb{R}^{2g}$ acts on $\mathcal{R}(\pi)$ freely and properly so that the orbit map becomes a principal bundle

$$\mathbb{R}^{2g} \to \mathcal{R}(\pi) \to \mathcal{R}(Q).$$

Since $\mathbb{R}^{2g}$ is contractible, its classifying space is a point and consequently $\mathcal{R}(\pi)$ splits as $(\mathcal{R}(\pi), \mathbb{R}^{2g}) = (\mathcal{R}(Q) \times \mathbb{R}^{2g}, \mathbb{R}^{2g})$ equivariantly, where $\mathbb{R}^{2g}$ acts on the second factor as translations.

Since $\mathbb{R}$ is the center of $\mathrm{Isom}_0(\widetilde{P})$, we have $\mu(\mathrm{Isom}_0(\widetilde{P})) = \mu(P)$ and

$$\mathcal{T}(\pi) = \mu(\mathrm{Isom}_0(\widetilde{P})) \setminus \mathcal{R}(\pi) = \mu(P) \setminus (\mathcal{R}(Q) \times \mathbb{R}^{2g}).$$

Now $\mu(P)$ acts on $\mathcal{R}(Q)$ freely and properly with quotient $\mathcal{T}(Q)$ [44] which has the homotopy type of the set of two points. Therefore $\mathcal{T}(\pi)$ is a product $\mathcal{T}(Q) \times \mathbb{R}^{2g}$. Moreover, it is known that $\mathcal{T}(Q)$ is diffeomorphic to two copies of $\mathbb{R}^{6g-6+2p}$, where p is the number of non-free orbit types of the Q action on $\mathbb{H}$. (This is the number of distinct conjugacy classes of maximal finite subgroups of Q.)

For the space of subgroups

$$\mathcal{S}(\pi) = \mathcal{R}(\pi)/\mathrm{Aut}(\pi) \cong (\mathcal{R}(Q) \times \mathbb{R}^{2g})/\mathrm{Aut}(\pi),$$

note that $\mathrm{Aut}(\pi) \cap \mathbb{R}^{2g} = \mathbb{Z}^{2g}$ and the quotient $\mathrm{Aut}(\pi)/\mathbb{Z}^{2g}$, which we called $\mathrm{Aut}(Q(\pi))$, is a subgroup of $\mathrm{Aut}(Q)$. By first dividing out by $\mathbb{Z}^{2g}$, we get $\mathcal{S}(\pi) = (\mathcal{R}(Q) \times T^{2g})/\mathrm{Aut}(Q(\pi))$. Since $\mathrm{Aut}(Q(\pi))$ acts on $\mathcal{R}(Q)$ freely, we have a genuine fibration

$$T^{2g} \longrightarrow \mathcal{S}(\pi) \longrightarrow \mathcal{S}(Q(\pi)).$$

The action of $\mathrm{Aut}(\pi)$ on $\mathbb{R}^{2g} \times \mathcal{R}(Q) = \mathcal{R}(\pi)$ is weakly $\mathbb{R}^{2g}$-equivariant, because $\mathbb{Z}^{2g} = \mathrm{Hom}(Q; \mathbb{Z})$ is normal in $\mathrm{Aut}(\pi)$. In other words,

$$\begin{aligned}\mathrm{Aut}(\pi) &\hookrightarrow \mathrm{TOP}_{\mathbb{R}^{2g}}(\mathbb{R}^{2g} \times \mathcal{R}(Q)) \\ &= \mathrm{M}(\mathcal{R}(Q), \mathbb{R}^{2g}) \rtimes (\mathrm{GL}(2g, \mathbb{R}) \times \mathrm{TOP}(\mathcal{R}(Q)))\end{aligned}$$

so that

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{Z}^{2g} & \longrightarrow & \mathrm{Aut}(\pi) & \longrightarrow & \mathrm{Aut}(Q(\pi)) & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \mathrm{M}(\mathcal{R}(Q), \mathbb{R}^{2g}) & \longrightarrow & \mathrm{TOP}_{\mathbb{R}^{2g}}(\mathbb{R}^{2g} \times \mathcal{R}(Q)) & \longrightarrow & \mathrm{GL}(2g, \mathbb{R}) \times \mathrm{TOP}(\mathcal{R}(Q)) & \longrightarrow & 1
\end{array}$$

is commutative. Thus, the structure group is a subgroup of the affine group of the torus $T^{2g} \circ \mathrm{GL}(2g, \mathbb{Z})$. Let $\theta \in \mathcal{R}(\pi)$. Then for $\alpha \in \pi$, $\theta \cdot \mu(\alpha) = \mu(\theta(\alpha)) \circ \theta$. Therefore, on $\mathcal{T}(\pi)$, $\mathrm{Inn}(\pi) = \mathrm{Inn}(\mathbb{R}\pi) = Q$ acts trivially as does $\mathrm{Inn}(Q) \cong Q$ on $\mathcal{T}(Q)$. Consequently, we have properly discontinuous actions of $\mathrm{Out}(\mathbb{R}\pi)$ and $\mathrm{Out}(Q(\pi))$ on $\mathcal{T}(\pi)$ and $\mathcal{T}(Q)$.

The space of moduli $\mathcal{M}(\pi) = \mathcal{T}(\pi)/\mathrm{Out}(\pi)$ requires more care. Recall that $\mathcal{R}(\pi) = \mathbb{R}^{2g} \times \mathcal{T}(Q)$ with $\mathbb{R}^{2g}$ action by translations on the first factor. Since

$$1 \longrightarrow \mathbb{R}^{2g} \longrightarrow \mathrm{Out}(\mathbb{R}\pi) \longrightarrow \mathrm{Out}(Q) \longrightarrow 1$$

is exact, we also have the commutative diagram:

$$\begin{array}{ccc}
(\mathcal{T}(\pi) = \mathbb{R}^{2g} \times \mathcal{T}(Q), \mathrm{Out}(\mathbb{R}\pi)) & \xrightarrow{/\mathbb{R}^{2g}} & (\mathcal{T}(Q), \mathrm{Out}(Q)) \\
\big\downarrow{\scriptstyle /\mathrm{Out}(\pi)} & & \big\downarrow{\scriptstyle /\mathrm{Out}(Q(\pi))} \\
\mathcal{M}(\pi) = \mathcal{T}(\pi)/\mathrm{Out}(\pi) & \xrightarrow{\quad q \quad} & \mathcal{M}(Q(\pi))
\end{array} \tag{6.6}$$

The actions and maps arise from the embedding

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \mathbb{Z}^{2g} & \longrightarrow & \mathrm{Out}(\pi) & \longrightarrow & \mathrm{Out}(Q(\pi)) & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
1 & \longrightarrow & \mathbb{R}^{2g} & \longrightarrow & \mathrm{Out}(\mathbb{R}\pi) & \longrightarrow & \mathrm{Out}(Q) & \longrightarrow & 1
\end{array} \tag{6.7}$$

obtained from Lemma 6.5.3 by dividing out the ineffective Q. Note $\mathrm{Inn}(\mathbb{R}\pi) \cap \mathbb{R}^{2g} = 1$. Now as $\mathrm{Out}(\pi)$ normalizes $\mathbb{R}^{2g}$ and $\mathbb{Z}^{2g}$ sits in $\mathbb{R}^{2g}$ as a lattice, the mapping q is a Seifert fibering with typical fiber the torus $\mathbb{R}^{2g}/\mathbb{Z}^{2g}$. In general, the fibering will not be locally trivial. In fact, if $F = (\mathrm{Out}(Q(\pi)))_{[\bar{\theta}]}$ for some $[\bar{\theta}] \in \mathcal{T}(Q)$, then the induced extension

$$1 \longrightarrow \mathbb{Z}^{2g} \longrightarrow E \longrightarrow F \longrightarrow 1$$

acts affinely on $\mathbb{R}^{2g} \times [\bar{\theta}]$ sitting over $[\bar{\theta}]$. The orbit over $[\bar{\theta}]$ under $\mathrm{Out}(Q(\pi))$ determines a 2-dimensional hyperbolic orbifold up to isometry. Over this hyperbolic orbifold is the set $E \setminus \mathbb{R}^{2g} = F \setminus T^{2g}$ of metric Seifert orbifolds in $\mathcal{M}(\pi)$ with base this hyperbolic orbifold. □

6.6. *Polynomial structures on virtually poly-$\mathbb{Z}$ manifolds*

We have seen in Theorem 5.2.3 that any torsion-free virtually poly-$\mathbb{Z}$ group π is the fundamental group of a closed $K(\pi, 1)$-manifold. For example, see (4.1). On the other hand, Milnor [46] has shown that such a group π can be the fundamental group of a complete affinely flat (not necessarily compact) manifold. The question arises as to whether one can find a *compact* complete affinely flat manifold with such fundamental group. It is true when π is 3-step nilpotent [57].

However, counter-examples produced by Benoist [5] and Burde and Grunewald [10] show that certain compact nilmanifolds do not admit a complete affinely flat structure. Consequently, the question has a negative answer even in the nilpotent case.

Note that one can look at such a complete affinely flat structure as a "polynomial structure of degree 1". In a situation where "polynomial structures of degree 1" fails to exist, the next best structure will be "polynomial structure of higher degree". The main reference for this section is [14].

A torsion-free filtration for a torsion-free finitely generated nilpotent group Γ is a central series of the form:

$$\Gamma_*\colon \Gamma_0 = 1 \subset \Gamma_1 \subset \Gamma_2 \subset \cdots \subset \Gamma_{c-1} \subset \Gamma_c \subset \Gamma_{c+1} = \Gamma$$

for which

$$\Gamma_{i+1}/\Gamma_i \cong \mathbb{Z}^k \quad \text{for } 1 \leqslant i \leqslant c \text{ and some } k \in \mathbb{N}_0.$$

Moreover, each Γ_i can be chosen to be a characteristic subgroup of Γ. We use K to denote the Hirsh number (or rank) of Γ. Often, we will also use $K_i = k_i + k_{i+1} + \cdots + k_c$. It follows that $K = K_1$.

NOTATION 6.6.1. (1) $P(\mathbb{R}^K, \mathbb{R}^k) \subset \mathrm{M}(\mathbb{R}^K, \mathbb{R}^k)$ denotes the vector space of polynomial mappings $p : \mathbb{R}^K \to \mathbb{R}^k$. So p is given by k polynomials in K variables.

(2) $P(\mathbb{R}^K) \subset \mathrm{TOP}(\mathbb{R}^K)$ will be used to indicate the set of all polynomial diffeomorphisms of $\mathbb{R}^K$, with an inverse which is also a polynomial mapping. This is a group where the multiplication is given by composition.

It is not hard to verify that $P(\mathbb{R}^K, \mathbb{R}^k)$ is a $\mathrm{Aut}(\mathbb{Z}^k) \times P(\mathbb{R}^K)$-module and that the resulting semi-direct product group

$$P(\mathbb{R}^K, \mathbb{R}^k) \rtimes \left(\mathrm{Aut}(\mathbb{Z}^k) \times P(\mathbb{R}^K)\right) \subseteq P\left(\mathbb{R}^{K+k}\right)$$

by defining $\forall (p, g, h) \in P(\mathbb{R}^K, \mathbb{R}^k) \rtimes (\mathrm{Aut}(\mathbb{Z}^k) \times P(\mathbb{R}^K))$, $\forall (x, y) \in \mathbb{R}^{k+K}$:

$$^{(p,g,h)}(x, y) = \left(gx - p\left(h(y)\right), h(y)\right).$$

Restrict $\mathrm{M}(\mathbb{R}^K, \mathbb{R}^k)$ to $P(\mathbb{R}^K, \mathbb{R}^k)$ and $\mathrm{TOP}(\mathbb{R}^K)$ to $P(\mathbb{R}^K)$, and we will speak of canonical type polynomial representations.

From now on, if we speak of a polynomial representation $\rho : \Gamma \to P(\mathbb{R}^K)$ which is of canonical type, we mean of canonical type with respect to some torsion-free central series.

THEOREM 6.6.2. (Existence) *Let Γ be a torsion-free, finitely generated nilpotent group of rank K. For any torsion-free central series Γ_*, there exists a canonical type polynomial representation $\rho : \Gamma \to P(\mathbb{R}^K)$.*

(Uniqueness) *If ρ_1, ρ_2 are two such polynomial representations of Γ into $P(\mathbb{R}^K)$ with respect to the same Γ_*, there exists a polynomial map $p \in P(\mathbb{R}^K)$ such that $\rho_2 = p^{-1} \circ \rho_1 \circ p$. Consequently, the manifolds $\rho_1(\Gamma)\backslash\mathbb{R}^K$ and $\rho_2(\Gamma)\backslash\mathbb{R}^K$ are "polynomially diffeomorphic".*

PROOF. The fundamental fact that we shall use is the cohomology vanishing

$$H^i\big(\Gamma/\Gamma_1, P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})\big) = 0$$

for $i > 0$. See [14] for a proof.

(Existence) We will proceed by induction on the nilpotency class c of Γ. If Γ is Abelian, then the existence is well known. Now, suppose that Γ is of class $c > 1$ and the existence is guaranteed for lower nilpotency classes. Using the induction hypotheses, the group Γ/Γ_1 can be furnished with a canonical type polynomial representation

$$\bar{\rho} : \Gamma/\Gamma_1 \to P(\mathbb{R}^{K_2}).$$

We obtain an embedding $i : \Gamma_1 \cong \mathbb{Z}^{k_1} \to P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})$ if we define $i(z) : \mathbb{R}^{K_2} \to \mathbb{R}^{k_1} : x \mapsto z$. We are looking for a map ρ making the following diagram commutative:

$$\begin{array}{ccccccccc}
1 & \longrightarrow & \Gamma_1 & \longrightarrow & \Gamma & \longrightarrow & \Gamma/\Gamma_1 & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle i} & & \downarrow{\scriptstyle \rho} & & \downarrow{\scriptstyle \varphi\times\bar{\rho}} & & \\
1 & \longrightarrow & P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1}) & \longrightarrow & P \rtimes (A \times P) & \longrightarrow & \mathrm{Aut}(\mathbb{Z}^{k_1}) \times P(\mathbb{R}^{K_2}) & \longrightarrow & 1 \\
 & & & & \cap & & & & \\
 & & & & P(\mathbb{R}^K) & & & &
\end{array}$$

where $P \rtimes (A \times P)$ denotes $P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1}) \rtimes (\mathrm{Aut}(\mathbb{Z}^{k_1}) \times P(\mathbb{R}^{K_2}))$ and $\varphi : \Gamma/\Gamma_1 \to \mathrm{Aut}(\mathbb{Z}^{k_1}) = \mathrm{Aut}(\Gamma_1)$ denotes the morphism induced by the extension $1 \to \Gamma_1 \to \Gamma \to \Gamma/\Gamma_1 \to 1$.

The existence of such a map ρ is now guaranteed by the surjectiveness of δ in the long exact cohomology sequence

$$\begin{aligned}
\cdots \longrightarrow 0 = H^1\big(\Gamma/\Gamma_1, P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})\big) &\longrightarrow H^1\big(\Gamma/\Gamma_1, P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})/\mathbb{Z}^{k_1}\big) \\
\xrightarrow{\delta} H^2\big(\Gamma/\Gamma_1, \mathbb{Z}^{k_1}\big) &\longrightarrow H^2\big(\Gamma/\Gamma_1, P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})\big) = 0 \longrightarrow \cdots.
\end{aligned}$$

(Uniqueness) Again we proceed by induction on the nilpotency class c of Γ. For $c = 1$ the result is again well known. Indeed, two canonical type representations of a virtually Abelian group are even known to be affinely conjugate.

So we suppose that Γ is of class $c > 1$ and that the theorem holds for smaller nilpotency classes. The representations ρ_1, ρ_2 induce two canonical type polynomial representations

$$\bar{\rho}_1, \bar{\rho}_2 : \Gamma/\Gamma_1 \to \mathbb{R}^{K_2}.$$

By the induction hypothesis, there exists a polynomial map $\bar{q} : \mathbb{R}^{K_2} \to \mathbb{R}^{K_2}$ such that

$$\bar{\rho}_2 = \bar{q}^{-1} \circ \bar{\rho}_1 \circ \bar{q}.$$

Lift this $\bar{q}$ to a polynomial map q of $\mathbb{R}^{K_1}$ as follows:

$$\forall x \in \mathbb{R}^{k_1}, \forall y \in \mathbb{R}^{K_2}: \; q : \mathbb{R}^{K_1} \to \mathbb{R}^{K_1} : (x, y) \mapsto q(x, y) = \big(x, \bar{q}(y)\big).$$

When restricted to Γ_1, ρ_1 and ρ_2 are mapping elements onto pure translations of $\mathbb{R}^{k_1}$. We know that there exist an affine mapping $\tilde{A} : \mathbb{R}^{k_1} \to \mathbb{R}^{k_1}$ for which

$$\rho_2|_{\mathbb{R}^{k_1}} = \tilde{A}^{-1} \circ \rho_1|_{\mathbb{R}^{k_1}} \circ \tilde{A}.$$

Extend $\tilde{A}$ to an affine mapping of $\mathbb{R}^K$ by defining

$$\forall x \in \mathbb{R}^{k_1}, \forall y \in \mathbb{R}^{K_2}: \; A : \mathbb{R}^{K_1} \to \mathbb{R}^{K_1} : (x, y) \mapsto A(x, y) = \big(\tilde{A}(x), y\big).$$

Let us denote $\psi = A^{-1} \circ q^{-1} \circ \rho_1 \circ q \circ A$. Then we see that ψ and ρ_2 are two canonical type polynomial representations of Γ, which coincide with each other on Γ_1 and which induce the same representation of Γ/Γ_1. This means that ψ and ρ_2 can be seen as the result of a Seifert construction with respect to the same data (cf. the commutative diagram above).

Now we use the injectiveness of δ, which implies that ψ and ρ_2 are conjugate to each other by an element $r \in P(\mathbb{R}^{K_2}, \mathbb{R}^{k_1})$ (seen as an element of $P(\mathbb{R}^K)$!). So we may conclude that

$$\rho_2 = r^{-1} \circ \psi \circ r = r^{-1} \circ A^{-1} \circ q^{-1} \circ \rho_1 \circ q \circ A \circ r = p^{-1} \circ \rho_1 \circ p$$

if we take $p = q \circ A \circ r$. □

References

[1] A. Adem and J.F. Davis, *Topics in transformation groups*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 1–54.

[2] L. Auslander, *Bieberbach's theorem on space groups and discrete uniform subgroups of Lie groups*, Ann. of Math. **71** (3) (1960), 579–590.

[3] L. Auslander, *On radicals of discrete subgroups of Lie groups*, Amer. J. Math. (1963), 145–150.

[4] L. Auslander and F.E.A. Johnson, *On a conjecture of C.T.C. Wall*, J. London Math. Soc. (2) **14** (1976), 331–332.

[5] Y. Benoist, *Une nilvariété non affine*, C. R. Acad. Sci. Paris Sér. I **315** (1992), 983–986.

[6] A. Borel et al., *Seminar on Transformation Groups*, Ann. of Math. Stud., Vol. 46, Princeton University Press (1960).
[7] G.E. Bredon, *Introduction to Compact Transformation Groups*, Academic Press, Inc., Orlando, FL (1972).
[8] W. Browder and W.C. Hsiang, *G-actions and the fundamental group*, Invent. Math. **65** (1982), 411–424.
[9] K.S. Brown, *Cohomology of Groups*, Grad. Texts in Math., Vol. 87, Springer, New York (1982).
[10] D. Burde and F. Grunewald, *Modules for certain Lie algebras of maximal class*, J. Pure Appl. Algebra **99** (1995), 239–254.
[11] H. Cartan and S. Eilenberg, *Homological Algebra*, Princeton University Press (1956).
[12] P.E. Conner and F. Raymond, *Actions of compact Lie groups on aspherical manifolds*, Topology of Manifolds, J.C. Cantrell and C.H. Edwards, eds, Markham, Chicago (1969), 227–264.
[13] P.E. Conner and F. Raymond, *Holomorphic Seifert fiberings*, Proc. 2nd Conference on Transformations groups, Part II, Lecture Notes in Math., Vol. 299 (1971), 124–204.
[14] K. Dekimpe, P. Igodt and K.B. Lee, *Polynomial structures for nilpotent groups*, Trans. Amer. Math. Soc. (1996), 77–97.
[15] H. Donnelly and R. Schultz, *Compact group actions and maps into aspherical manifolds*, Topology **21** (1982), 443–455.
[16] G. Dula and D. Gottlieb, *Splitting off H-spaces and Conner–Raymond splitting theorem*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **37** (2) (1990), 321–334.
[17] F.T. Farrell and W.C. Hsiang, *Topological characterization of flat and almost flat Riemannian manifolds* M^n $(n \neq 3, 4)$, Amer. J. Math. **105** (1983), 641–672.
[18] W. Goldman, *Nonstandard Lorentz space forms*, J. Differential Geom. **21** (1985), 301–308.
[19] D. Gottlieb, *Splitting off tori and the evaluation subgroup of the fundamental group*, Israel J. Math. (1989), 216–222.
[20] D. Gottlieb, K.B. Lee and M. Özaydın, *Compact group actions and maps into* $k(\pi, 1)$*-spaces*, Trans. Amer. Math. Soc. (1985), 419–429.
[21] H. Holmann, *Seifertsche Faseraume*, Math. Ann. **157** (1964), 138–166.
[22] P. Igodt, *Generalizing a realization result of B. Zimmermann and K.B. Lee to infranilmanifolds*, Manuscripta Math. (1984), 19–30.
[23] S. Kerckhoff, *The Nielsen realization problem*, Ann. of Math. (2) **117** (1983), 235–265.
[24] F.W. Kamber and Ph. Tondeur, *Flat manifolds with parallel torsion*, J. Differential Geom. **2** (1968), 385–389.
[25] E. Kang and K.B. Lee, *Injective operations of homogeneous spaces*, Michigan Math. J. (1997), to appear.
[26] Y. Kamishima, K.B. Lee and F. Raymond, *The Seifert construction and its applications to infranilmanifolds*, Quart. J. Math. Oxford **34** (1983), 433–452.
[27] H. Kim, *Complete left-invariant affine structures on nilpotent Lie groups*, J. Differential Geom. (1986), 373–394.
[28] H.T. Ku and M.C. Ku, *Group actions on aspherical* $A_k(N)$*-manifolds*, Trans. Amer. Math. Soc. **278** (1983), 841–859.
[29] R. Kulkarni, K.B. Lee and F. Raymond, *Deformation spaces for Seifert manifolds*, Geometry and Topology, University of Maryland 1983–1984, Vol. 1167, J. Alexander and J. Harer, eds (1986), 180–216.
[30] R. Kulkarni and F. Raymond, *3-dimensional Lorentz space-forms and Seifert fiber spaces*, J. Differential Geom. (1985), 231–268.
[31] J.B. Lee, K.B. Lee and F. Raymond, *Seifert manifolds with fiber spherical space forms*, Trans. Amer. Math. Soc. **348** (1996), 3763–3798.
[32] K.B. Lee, *Geometric realization of* $\pi_0 \varepsilon(m)$, Proc. Amer. Math. Soc. **86** (2) (1982), 353–357.
[33] K.B. Lee, *Geometric realization of* $\pi_0 \varepsilon(m)$ *II*, Proc. Amer. Math. Soc. **87** (1982), 175–178.
[34] K.B. Lee, *Aspherical manifolds with virtually 3-step nilpotent fundamental group*, Amer. J. Math. **105** (1983), 1435–1453.
[35] K.B. Lee, *Infra-solvmanifolds of type (R)*, Quart. J. Math. Oxford (2) **43** (1995), 185–195.
[36] K.B. Lee and F. Raymond, *Topological, affine and isometric actions on flat Riemannian manifolds*, J. Differential Geom. **16** (1981), 255–269.
[37] K.B. Lee and F. Raymond, *Topological, affine and isometric actions on flat Riemannian manifolds II*, Topology Appl. **13** (1982), 295–310.

[38] K.B. Lee and F. Raymond, *Geometric realization of group extensions by the Seifert construction*, Contemp. Math. **33** (1984), 353–411.

[39] K.B. Lee and F. Raymond, *Rigidity of almost crystallographic groups*, Contemp. Math. **44** (1985), 73–78.

[40] K.B. Lee and F.A. Raymond, *Manifolds on which only tori can act*, Trans. Amer. Math. Soc. (1987), 487–499.

[41] K.B. Lee and F. Raymond, *Seifert manifolds modelled on principal bundles*, Transf. Groups (Osaka, 1987), Lectures Notes in Math., Vol. 1375 (1989), 207–215.

[42] K.B. Lee and F. Raymond, *Seifert manifolds with $\Gamma\backslash G/K$-fiber*, Michigan Math. J **43** (1996), 437–464.

[43] S. MacLane, *Homology*, Grundlehren Math. Wiss., Vol. 114, Springer, Berlin (1975).

[44] A.M. Macbeath and D. Singleman, *Spaces of subgroups and Teichmüller spaces*, Proc. London Math. Soc. (3) (1975), 211–256.

[45] A. Mal'cev, *On a class of homogeneous spaces*, Trans. Amer. Math. Soc. **1** (9) (1962), 276–307.

[46] J. Milnor, *On fundamental groups of complete affinely flat manifolds*, Adv. Math. **25** (1977), 178–187.

[47] A. Nicas and C. Stark, *K-theory and surgery of codimension-two torus actions on aspherical manifolds*, J. London Math. Soc. **31** (2) (1985), 173–183.

[48] P. Orlik, *Seifert Manifolds*, Lecture Notes in Math., Vol. 291, Springer (1972).

[49] R. Palais, *On the existence of slices for actions of non-compact Lie groups*, Ann. of Math (1961), 295–323.

[50] C.Y. Park, *Homotopy groups of automorphisms of some injective Seifert fiber spaces*, Thesis, University of Michigan (1989).

[51] M.S. Raghunathan, *Discrete Subgroups of Lie Groups*, Ergeb. Math. Grenzgeb., Vol. 68, Springer (1972).

[52] F. Raymond, *The Nielsen theorem for Seifert fiber spaces over locally symmetric spaces*, J. Korean Math. Soc. **16** (1979), 87–93.

[53] F. Raymond and L. Scott, *The failure of Nielsen's theorem in higher dimensions*, Arch. Math. **29** (1977), 643–654.

[54] F. Raymond and D. Wigner, *Construction of aspherical manifolds*, Geom. Appl. Homotopy Theory **657** (1977), 408–422.

[55] F.A. Raymond and A. Vasquez, *3-Manifolds whose universal coverings are Lie groups*, Topology Appl. **12** (1981), 161–179.

[56] H. Rees, *Special manifolds and Novikov's conjecture*, Topology **22** (1983), 365–378.

[57] J. Scheuneman, *Affine structures on three-step nilpotent Lie algebras*, Proc. Amer. Math. Soc. **46** (3) (1974), 451–454.

[58] P. Scott, *The geometries of 3-manifolds*, Bull. London Math. Soc. **15** (1983), 401–487.

[59] R. Schoen and S.T. Yau, *Compact group actions and the topology of manifolds with non-positive curvature*, Topology **18** (1979), 361–380.

[60] H. Seifert, *Topologie drei-dimensionaler gefaserter Räume*, Acta. Math. **60** (1933), 147–238.

[61] F. Waldhausen, *Gruppen mit Zentrum und drei-dimensionaler Mannigfaltigkeiten*, Topology **6** (1967), 505–517.

[62] R. Washiyama and T. Watabe, *On the degree of symmetry of a certain manifold*, J. Math. Soc. Japan **35** (1983), 53–58.

[63] H. Zieschang, *On toric fiberings over surfaces*, Math. Notes (1969), 341–345.

[64] H. Zieschang and B. Zimmermann, *Endliche Gruppen von Abbildungsklassen gefaserter 3-Mannigfaltigkeiten*, Math. Ann. (1979), 41–62.

[65] B. Zimmermann, *Über Gruppen von Homoomorphismen Seifertscher Faserraume und flacher Mannigfaltigkeiten*, Manuscripta Math. **30** (1980), 361–373.

CHAPTER 14

Quantum Invariants of 3-Manifolds

W.B. Raymond Lickorish

Department of Pure Mathematics and Mathematical Statistics, C.M.S., Wilberforce Road, Cambridge, CB3 0WB, England, UK

E-mail: wbrl@dpmms.cam.ac.uk

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

The theory of quantum invariants of 3-manifolds is in a strange unsatisfactory state. On the one hand that makes it exciting and yet it is also somewhat bewildering. Descriptions of the invariants sometimes require considerable prior knowledge in some speciality. The invariants obtained by different routes tend to differ, by various normalisations and constants, and that makes translation arduous. Very many people are working in and around this subject. For many of them 3-manifold theory itself is not really a major interest, for these invariants can be interpreted from view-points in Quantum Field Theory, Statistical Mechanics, Category Theory, Differential Geometry, Operator Algebras and probably several more. Within the ambiance of these invariants the new subjects of Quantum Groups [10,43] and Topological Quantum Field Theory [1] have been born; a recent survey is in [42]. This article tries to introduce one version of some of these quantum invariants to mathematicians who already have an interest in low-dimensional topology. It makes no attempt to unify vast areas of human knowledge, to search for ever deeper truth nor to develop the most sophisticated notations. The main emphasis will be on proving the existence of some non-trivial quantum invariants using the linear skein theory associated to the Kauffman bracket; this method was developed progressively in [23–25,27]. That method has the advantages of being fairly quick, self-contained and visualisable. A few calculations of the invariants will be made together with intimations of how other calculations might be made. It should be admitted, at the start, that, from the perspective of the established theory of 3-manifolds, these invariants have been disappointing. Applications are few and far between. Certainly the invariants do distinguish apart various pairs of 3-manifolds but such pairs were already known to be distinct. Sadly, the invariants fail to classify lens spaces. The invariants that will be described here are in some sense generalisations of the original Jones polynomial. That polynomial invariant has had great success in giving information about ways of drawing diagrams of knots. To find some similar relevance of the quantum invariants to any facet of standard 3-manifold theory has become a provocative challenge to low-dimensional topologists.

It is instructive to start at the beginning. According to legend, whilst dining with Sir Michael Atiyah in Annie's Café in Swansea, Ed Witten produced the following definition.

$$Z_k(M) = \int_{A\in\mathcal{A}} \mathrm{e}^{\frac{ik}{4\pi}\int_M \mathrm{tr}(\mathrm{d}A\wedge A+\frac{2}{3}A\wedge A\wedge A)}\,\mathrm{d}A.$$

This was to be the definition of an invariant $Z_k(M)$ of a closed orientable 3-manifold M, the invariant depending upon a 'level' k and a Lie group G with Lie algebra $\mathfrak{g}$; $\mathcal{A}$ denotes the moduli space of connections on a G-bundle over M. The expression $\int_M \mathrm{tr}(\mathrm{d}A \wedge A + \frac{2}{3}A \wedge A \wedge A)$ is a multiple of the famous Chern–Simons action on A, where the connection A is now regarded as a $\mathfrak{g}$-valued 1-form. If the main integrand were multiplied by a monodromy term

$$\prod_{i=1}^{n} \mathrm{tr}_{\rho_i}\left(\exp\left(\int_{L_i} A\right)\right)$$

this would become an invariant of a framed link L, with components $L_1, L_2, \dots, L_n$, contained in M with each L_i equipped with an irreducible representation ρ_i of $\mathfrak{g}$. The difficulty however is that the main integral over $\mathcal{A}$ has never been satisfactorily defined in spite of much effort. Nevertheless formal manipulations of this integral have led to results later rigorously established. In particular, when M is the 3-sphere S^3, $G = \mathrm{SU}(2)$ and L is a 0-framed link equipped with the 2-dimensional representation on every component, then the invariant becomes the Jones polynomial $V_L(t)$ evaluated at $\exp(\frac{2\pi i}{k+2})$. Furthermore, the invariant should interact well with the idea of surgery. The first proofs that invariants exist in accord with these ideas were given given by Reshetikhin and Turaev [39] for SU(2) and Turaev and Wenzl [45] for other Lie groups using representations of universal enveloping algebras with certain roots of unity for their parameter q.

2. The Kirby calculus

All proofs of the existence of quantum invariants seem to start with a 3-manifold given in terms of surgery on a framed link in S^3. A framed link L in S^3 is a link together with a never-zero cross-section of its normal bundle. Thus L can be thought of as a link of bands (long thin annuli), two such links being equivalent if the bands of one can be moved to those of the other (by ambient isotopy). If one orients the two boundary components of a band, in the same direction around the band, and considers their linking number, that gives an integer *framing number* that specifies the framing of the component considered. The framing also specifies up to isotopy a simple closed curve (a boundary component of the band) on the boundary of a tubular neighbourhood $N(L_i)$ of each component L_i. That is called the framing curve of L_i.

DEFINITION. The oriented 3-manifold M obtained from S^3 by surgery on the framed link L is given by

$$M = \overline{S^3 - N(L)} \cup (\text{solid tori}),$$

where one solid torus is glued by its boundary to each $\partial N(L_i)$ in such a way that a curve bounding a disc in the solid torus is identified with the framing curve.

A classical result about 3-manifolds is, then, the following [21,46].

THEOREM 2.1. *Any closed connected oriented 3-manifold can be obtained from S^3 by surgery on some framed link.*

It is easy to give examples [22] in which surgeries on different framed knots gives the same 3-manifold, and even easier to do this for links. It is the next result of Kirby [12] that explains necessary and sufficient conditions for this to happen.

THEOREM 2.2. *Two framed links specify by surgery the same 3-manifold, up to orientation preserving homeomorphism, if they differ by (ambient) isotopy in S^3 and a sequence of moves of the following two types.*

Fig. 1.

K(1) *Add or subtract an unknot with framing* ± 1 *that is contained in a ball otherwise disjoint from the links.*

K(2) *Suppose two components of a framed link and their framing annuli lie, as in one of the diagrams of Figure* 1, *in a doubly punctured disc that is contained in* S^3 *(in maybe a very knotted, twisted, linked-up way). Replace those two framed components by the two in the other diagram of Figure* 1 *in the same doubly punctured disc.*

This theorem is often viewed in terms of adding and sliding 2-handles attached to a 4-ball, the boundary of the whole ensemble being the given 3-manifold. The proof of the theorem in [12] is an application of Cerf Theory to such 4-manifolds.

These two results mean that, to define a 3-manifold invariant, it is just necessary to associate some mathematical entity to a framed link and then prove it invariant under moves of types K(1) and K(2). As an easy example, orient a framed link L and consider the matrix Λ of linking numbers, $\Lambda_{i,j} = \text{lk}(L_i, L_j)$, with $\Lambda_{i,i}$ being the framing number of the component L_i. It is easy to see that the Abelian group presented by this matrix of integers is invariant under the K-moves. However this group is easily shown to be the first homology group of the 3-manifold obtained by surgery on L.

The idea of what follows is to use some polynomial invariant of the framed link L. It is not quite that easy however. What does work is a linear combination of the Jones polynomials, evaluated at certain complex roots of unity, of links formed by adding various parallels to the components of L. That, together with an adjustment depending only on the signature of the above linking matrix Λ will give a 3-manifold invariant. The complications, needed in defining the coefficients in the linear combination, are eased by the systematic approach of linear skein theory which will now be described.

3. Skein theory based on the Kauffman bracket

Let M be an oriented 3-manifold with a finite set (maybe empty) of framed points in its boundary. The framing of such a point x is a specified element of the tangent space of ∂M at x. Let A be a fixed complex number; later it will be a root of unity.

DEFINITION. The linear skein space $\mathcal{S}M$ is the vector space over $\mathbb{C}$ generated by all links of framed simple closed curves and framed proper arcs in M, that meet ∂M in precisely the given framed points, quotiented by the three relations of Figure 2.

Here (i) means that the union of L with a 0-framed unknot in a ball disjoint from L is equal to $(-A^{-2} - A^2)L$. The equality of (ii) refers to three framed links that are identical

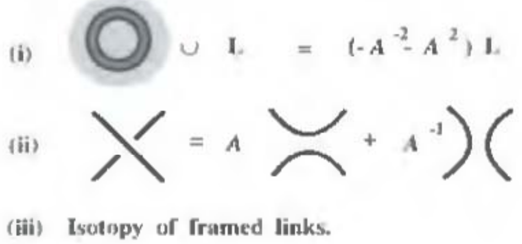


Fig. 2.

Fig. 3.

except in a ball in which they are as shown, all framing directions pointing towards the eye of the reader. As an exercise prove that, if a framed link L is changed to L' by adding one complete right-hand twist to the framing of one component, then $L' = -A^3 L$ in $\mathcal{S}M$. Elements of $\mathcal{S}M$ are sometimes called skeins in M.

First consider $\mathcal{S}S^3$. All links can be projected to become link diagrams in S^2 with the framings projecting to be the normal to S^2. In this way a link diagram represents a *framed* link. Ambient isotopy of framed links in S^3 translates to equivalence of diagrams by Reidemeister moves of types II and III, see Figure 3 (and homeomorphisms of S^2).

A few moments thought will show that quotienting by (i) and (ii) implies the quotienting by these two types of Reidemeister move. Thus $\mathcal{S}S^3$ is isomorphic to the space spanned by link diagrams quotiented by (i) and (ii). Then very simple induction arguments concerning the number of crossings in a diagram show that $\mathcal{S}S^3$ is a 1-dimensional vector space; the empty link will be taken to be a natural base. To find the coordinate of a framed link, represent it by a diagram with n crossings, use (ii) to express it as a sum of 2^n diagrams with no crossing, then in each diagram remove all the resulting curves with no crossing by using (i). This is, of course, the essence of the Kauffman bracket approach to the Jones polynomial. If the 0-framed unknot were taken as base, the coordinate of a 0-framed knot would be that knot's Jones polynomial in $t = A^{-4}$ (where A is here being regarded as an indeterminate, not a complex number).

The same method applies to the product of any surface F with an interval. The skein space has a base of framed links represented by all link diagrams in F that have no crossing and no null-homotopic curve. Thus a solid torus, parametrised as $S^1 \times D^2$, projects to the annulus $S^1 \times I$ where I is a diameter of D^2. Then $\mathcal{S}(S^1 \times D^2)$ has, as a base, framed links represented by diagrams of the form $S^1 \times X \subset S^1 \times I$ where X is a finite set. Now take a pair of disjoint orientation-preserving embeddings of D^2 into D^2 and take the product

Fig. 4.

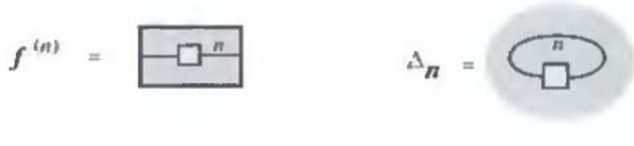


Fig. 5.

with the identity on S^1 to obtain an embedding

$$(S^1 \times D^2) \sqcup (S^1 \times D^2) \to (S^1 \times D^2),$$

where '$\sqcup$' denotes disjoint union. This induces, by embedding pairs of framed links, a bilinear map

$$\mathcal{S}(S^1 \times D^2) \times \mathcal{S}(S^1 \times D^2) \to \mathcal{S}(S^1 \times D^2).$$

Under this bilinear map $\mathcal{S}(S^1 \times D^2)$ becomes a commutative algebra with just one generator, which will be called α, the framed link of one component represented by $(S^1 \times \text{one point}) \subset S^1 \times I$. Thus $\mathcal{S}(S^1 \times D^2)$ is the polynomial algebra $\mathbb{C}[\alpha]$.

A more subtle example of this skein theory is that of the 3-ball with $2n$ specified framed boundary points. That will be considered in terms of projections to link diagrams of arcs and closed curves in a rectangle, with n points on its left side and n on its right side and all framing directions pointing towards the viewer. As before that space has a base consisting of all diagrams of n arcs, with no crossing, joining up the $2n$ boundary points. The dimension of the space is the Catalan number $\frac{1}{n+1}\binom{2n}{n}$. However, juxtaposition of rectangles, identifying the right side of one with the left side of the next, at once induces a product on the space that makes it a (non-commutative) algebra. This is TL_n, the n-th Temperley–Lieb algebra. As an algebra it is generated by just the n elements, $\{1, e_1, e_2, \ldots, e_{n-1}\}$ that are shown in Figure 4. That diagram introduces the convention, to be used throughout, that an integer n beside an arc means that n parallel copies of the arc are present. Of course, 1 is a multiplicative identity of the algebra.

Now will be defined, by induction on n, the all important Jones–Wenzl (see [47]) idempotent $f^{(n)} \in TL_n$. It is characterised in Lemma 3.1. In diagrams that ensue, a small white square will represent this $f^{(n)}$; the number of arcs entering that square determines to which n it refers.

Suppose (the linear sum of diagrams that represents) $f^{(n)}$ is placed in S^2 and its $2n$ specified boundary points are joined up by n parallel standard arcs, as in Figure 5. The resulting element of $\mathcal{S}(S^3) \equiv \mathbb{C}$ so represented will be denoted Δ_n.

Fig. 6.

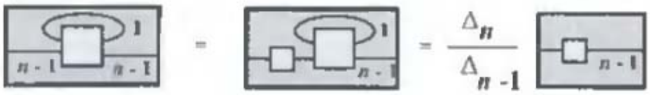


Fig. 7.

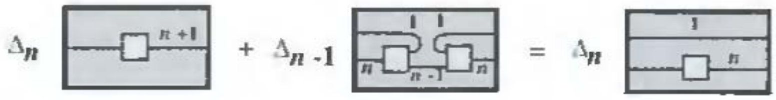


Fig. 8.

LEMMA 3.1. *Suppose that A^4 is not a k-th root of unity for $k \leqslant n$. Then there is a unique element $f^{(n)} \in TL_n$ such that*

(i) $f^{(n)}e_i = 0 = e_i f^{(n)}$ *for* $1 \leqslant i \leqslant n-1$,

(ii) $(f^{(n)} - 1)$ *belongs to the algebra generated by* $\{e_1, e_2, \dots, e_{n-1}\}$,

(iii) $f^{(n)} f^{(n)} = f^{(n)}$ *and*

(iv) $\Delta_n = \frac{(-1)^n(A^{2(n+1)} - A^{-2(n+1)})}{A^2 - A^{-2}}$.

PROOF. Uniqueness follows from (i) and (ii) as $1 - f^{(n)}$ is the identity of the sub-algebra generated by $\{e_1, e_2, \dots, e_{n-1}\}$. Further, (i) and (ii) imply (iii). Define $f^{(1)} = 1$, and $f^{(0)}$ to be the empty diagram and $\Delta_0 = 1$. Assume $f^{(k)}$ has been defined for $k \leqslant n$. The identity of Figure 6 follows for $i + j \leqslant n$ as $f^{(i)}$ is 1 plus a sum of products of the e_i. The latter terms die on contact with $f^{(i+j)}$.

Figure 7 shows an equality in TL_{n-1}. The second diagram is in $f^{(n-1)}TL_{n-1}$ and so (by (i)) is a multiple of $f^{(n-1)}$. The multiplier is determined by inserting both sides of the equation into copies of S^2 and joining up the specified boundary points in the standard way.

Suppose A^4 is not a k-th root of unity for $k \leqslant n+1$. In Figure 8 is the definition of $f^{(n+1)}$, valid since $\Delta_n \neq 0$ by (iv). Properties (i) and (ii) for $f^{(n+1)}$ then follow. All are immediate except for the property $f^{(n+1)}e_n = 0$. However, multiplying the second diagram of Figure 8 by e_n produces a diagram containing a copy of the first diagram of Figure 7. That can then be excised at the cost of multiplying by Δ_n/Δ_{n-1} (use Figure 6). There results the right-hand side of Figure 8 multiplied by e_n so $f^{(n+1)}e_n = 0$.

Now place this defining equation for $f^{(n+1)}$ into a disc with two holes (to be regarded as a projection of a genus 2 handlebody), joining the top two boundary points of the square around the left hole, and joining the other n points on the left to the other n on the right

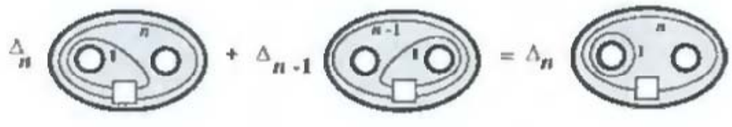


Fig. 9.

by standard arcs that encircle both holes. The result is shown in Figure 9. In the first and third diagrams that should be clear. To form the second diagram, the left-hand square representing $f^{(n)}$ in the second diagram of Figure 8 has been slid around near the boundary of the disc until it coalesces with the other copy of $f^{(n)}$ (using $f^{(n)}f^{(n)} = f^{(n)}$).

If discs are added to the boundary components of this twice punctured disc to give S^2, this becomes, on dividing by Δ_n, the equation

$$\Delta_{n+1} + \Delta_{n-1} = (-A^{-2} - A^2)\Delta_n.$$

The required formula (iv) for Δ_{n+1} follows from this at once by induction. □

Now suppose the right-hand hole of Figure 9 is filled with a disc. One obtains an equality in the annulus and so in the skein space of $S^1 \times D^2$. The first term is $f^{(n+1)}$ inserted into the annulus with points on its left joined to those on its right by standard arcs around the annulus. That must represent some element in the skein space of $\mathcal{S}(S^1 \times D^2) = \mathbb{C}[\alpha]$. Let it be denoted $S_{n+1}(\alpha)$. Then Figure 9, with help from Figure 7, gives

$$S_{n+1}(\alpha) = \alpha S_n(\alpha) - S_{n-1}(\alpha).$$

Together with $S_0(\alpha) = 1$ and $S_1(\alpha) = \alpha$, that is the defining recurrence relation for the Chebyshev polynomials of the second kind (renormalised).

DEFINITION. In the skein space of the solid torus, for each integer $r \geqslant 3$, define an element ω_r by

$$\omega_r = \sum_{n=0}^{r-2} \Delta_n S_n(\alpha) \in \mathcal{S}(S^1 \times D^2).$$

Now the left-hand side of Figure 9 is symmetric with respect to the two holes except for the allocation of n and $n-1$. Thus summing, from $n = 0$ to $n = r-2$, the difference between that identity and its analogue with the roles of the holes interchanged, produces the following.

LEMMA 3.2. *The difference between the skeins of Figure* 10 *can be expressed as the difference between two skeins, each containing a copy of* $f^{(r-1)}$.

Fig. 10.

Note that in the above, ω_r beside a framed curve really indicates the image of ω_r under a framing preserving embedding of $S^1 \times D^2$ into the handlebody. The handlebody contains the other framed curve as shown.

Now suppose that A is a primitive $4r$-th root of unity. Suppose that a sphere separates S^3 into two balls, one inside the sphere and the other outside it. Suppose that $f^{(r-1)} \in TL_{r-1}$ is placed in the inside ball and any element whatsoever, ξ, of TL_{r-1} in the outside ball (with framed points agreeing on the sphere). The resulting element of $\mathcal{S}S^3$ is zero. That is because ξ is a linear sum of products of the e_i and of the element $1 \in TL_{r-1}$. Of course $f^{(r-1)}e_i = 0$ and combining the $f^{(r-1)}$ in the inner ball with 1 in the outer ball gives Δ_{r-1}. For A a $4r$-th root of unity Δ_{r-1} is zero. That gives impetus to the following definition.

DEFINITION. Two elements of the skein space of a 3-manifold M are equal modulo $f^{(r-1)}$ if their difference is in the subspace of elements that consist of $f^{(r-1)}$ in a ball B in M together with any skein in $\overline{M - B}$ (with $2(r-1)$ specified points on ∂B).

The idea of skein space elements being equal modulo any other element of the skein of a ball, with boundary points, is defined in the same way. Theorem 3.1 reformulates the above remarks.

THEOREM 3.1. *The two elements of the skein of a genus two handlebody H shown in Figure 10 are equal modulo $f^{(r-1)}$. Any inclusion of H into S^3 induces a natural bilinear map*

$$\mathcal{S}H \times \mathcal{S}(S^3 - H) \to \mathcal{S}S^3$$

and, when A is a primitive $4r$-th root of unity, the two given elements produce the same element of the dual of $\mathcal{S}(S^3 - H)$.

The curves labelled 1 in Figure 10 can be replaced by any collection of curves parallel in the given framing (apply the theorem to them one at a time) and then, using bilinearity, by linear sums of such parallels. In particular, the curves labelled 1, referred to in Theorem 3.1, can be replaced by curves labelled ω_r. (Technically, 'curves labelled ω_r' means images of ω_r under inclusions of solid tori respecting framings.)

Suppose that L is a framed link of n components temporarily oriented and ordered. There are canonical embeddings up to isotopy of $S^1 \times D^2$ onto neighbourhoods of the components sending the natural product framing of $S^1 \times D^2$ to the given framings. This induces a multilinear map

$$\langle\ ,\ ,\dots,\rangle_L : \mathcal{S}(S^1 \times D^2) \times \mathcal{S}(S^1 \times D^2) \times \cdots \times \mathcal{S}(S^1 \times D^2) \to \mathcal{S}S^3.$$

Noting the similarity of Figures 10 and 1, the consequence of the above is this.

THEOREM 3.2. *When A is a primitive $4r$-th root of unity, the complex number $\langle \omega_r, \omega_r, \ldots, \omega_r \rangle_L$ does not change when L is changed by a* K(2) *move.*

4. Invariants of 3-manifolds from skein theory

The last result of the last section describes a way of associating a complex number to a framed link that is invariant under the second type of Kirby move on the link. It is now almost a matter of administration to produce an invariant for closed 3-manifolds. Suppose then that a 3-manifold M is obtained from S^3 by surgery on a framed link L and let b_+ and b_- be the number of positive and the number of negative eigenvalues of Λ, the linking matrix for L. It is easy to check that those numbers do not change under a K(2) move. Let U_+ and U_- be the unknot with framings ± 1.

THEOREM 4.1. *Let A be a primitive $4r$-th root of unity. The complex number*

$$\langle \omega_r, \omega_r, \ldots, \omega_r \rangle_L \langle \omega_r \rangle_{U_+}^{-b_+} \langle \omega_r \rangle_{U_-}^{-b_-}$$

is an invariant of M.

That follows at once from Theorem 3.2 and the Kirby moves, though it is just necessary also to check (see below) that $\langle \omega_r \rangle_{U_+} \langle \omega_r \rangle_{U_-} \neq 0$. Modulo possible re-normalisation, these invariants are, for $r \geqslant 3$, quantum $\mathrm{SU}_q(2)$ invariants of M.

The two identities of Figure 11 are exercises in the use of $f^{(n)}$. The first diagram refers to n parallel strings containing $f^{(n)}$ encircled by a parallel strings containing $f^{(a)}$ (that is, encircled by a copy of $S_a(\alpha)$). Prove the result from the definition of $f^{(n)}$ when $a = 1$ and recall the definition of the Chebyshev polynomials. The second formula is proved by induction on n.

Suppose that $A^{2n} \neq 1$ and $A^{2n+4} \neq 1$; that is certainly the case when A is a primitive $4r$-th root of unity. Consider the element x, say, of TL_n shown in Figure 12 that consists of $f^{(n)}$ encircled by ω_r. Modulo $f^{(r-1)}$, this x is *zero* when $n \neq 0$ and is $\langle \omega_r \rangle_U$ when $n = 0$ (here U is the 0-framed unknot). This is a most important fact which is proved in the following way. Note that $-(A^{-2} + A^2)x = -(A^{-2(n+1)} + A^{2(n+1)})x$, as is seen

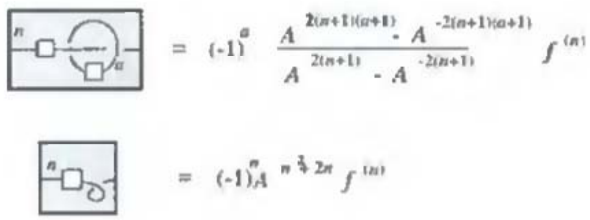


Fig. 11.

Fig. 12.

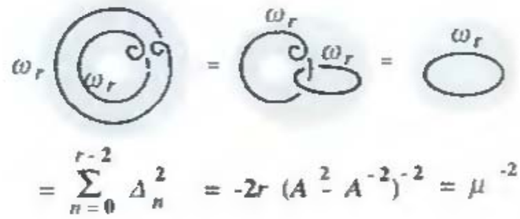


Fig. 13.

by introducing a small 0-framed unknot, slipping it over the component labelled ω_r (by Lemma 3.2) and then removing it using Figure 11 (when $a = 1$). The result follows at once.

That result together with the sliding property of ω_r shows, as in Figure 13, that $\langle\omega_r\rangle_{U_+}\langle\omega_r\rangle_{U_-} = \langle\omega_r\rangle_U$ when A is a primitive $4r$-th root of unity. However, use of the definitions of ω_r and of Δ_n (and summation of a geometric progression) shows that

$$\langle\omega_r\rangle_U = \sum_{n=0}^{r-2} \Delta_n^2 = -2r\left(A^2 - A^{-2}\right)^{-2}.$$

This is certainly non-zero and that tidies away the final detail of the proof of Theorem 4.1.

A little calculation shows that

$$\langle\omega_r\rangle_{U_+} = \sum_{n=1}^{4r} A^{n^2} / \left(2A^{3+r^2}\left(A^2 - A^{-2}\right)\right).$$

That, together with the formulae of Figure 11 gives enough information to calculate the invariants for lens spaces, Seifert fibrings over S^2 and, in fact, any 3-manifold with a surgery link consisting of N framed unknots simply linked together according to a pattern defined by a tree. A copy of ω_r must be placed on each of the N components, each ω_r must be expanded as $\sum_{n=0}^{r-2} \Delta_n S_n(\alpha)$ and then the value in $\mathcal{S}S^3$ of each of the resulting $(r-1)^N$ links (with components labelled with the $S_n(\alpha)$) must be methodically calculated using Figure 11, starting from a 'leaf' of the tree. There results a complicated summation. At least for lens spaces this has been extensively studied in [6,8,36], and [20] (watch for variants of notation and normalisation). It transpires that lens spaces L_{p,q_1} and L_{p,q_2} have the same $\mathrm{SU}_q(2)$ invariants for all r if $q_1^2 \equiv q_2^2 \equiv -1 \bmod p$ (for example, $L_{65,8}$ and $L_{65,18}$). Further, calculation shows that the homotopy equivalent 3-manifolds $L_{7,1}$ and $L_{7,2}$ have distinct $\mathrm{SU}_q(2)$ invariants.

It is easy to construct other pairs of 3-manifolds with all the same $\mathrm{SU}_q(2)$ invariants. Let X_1 and X_2 be the exteriors of two knots. Let $h : \partial X_1 \to \partial X_2$ be a homeomorphism and let $-h$ be that homeomorphism composed with the automorphism of ∂X_1 that induces minus the identity map on the first homology group. Then manifolds $X_1 \cup_h X_2$ and $X_1 \cup_{-h} X_2$ have all the same $\mathrm{SU}_q(2)$ invariants [9,26].

The ideas outlined above have been extended in various ways. In particular if M has a spin structure, Θ, an invariant of (M, Θ) can be defined so that the above invariant is, when r is even, the sum of these invariants over all spin structures. If M has a designated element of $H^1(M; \mathbb{Z}/2\mathbb{Z})$ then an invariant of the pair can also be defined. For details see [3,14] and [27]. An extension defining an invariant when A is a primitive $2r$-th root of unity, r being odd, was made in [4].

It is often convenient to normalise these invariants in a way that is slightly different from that of the above theorem. If throughout ω_r is multiplied by a constant complex number μ, another invariant is produced that is the first invariant multiplied by μ^{β_1} where β_1 is the first Betti number of the manifold M. Choose μ to be a real number such $\mu^2 = (A^2 - A^{-2})^2/(-2r)$ then the normalised invariant $\mathcal{I}_A(M)$ is defined by

$$\mathcal{I}_A(M) = \mu\langle \mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r\rangle_L \langle \mu\omega_r\rangle_{U_-}^{\sigma\Lambda},$$

where $\sigma\Lambda$ is the signature of the matrix of linking numbers of L. The first μ is a convention to achieve $\mathcal{I}_A(S^1 \times S^2) = 1$; it also gives $\mathcal{I}_A(S^3) = \mu$.

LEMMA 4.1. *If $\overline{M}$ is M with the other orientation then $\mathcal{I}_A(\overline{M}) = \overline{\mathcal{I}_A(M)}$. For connected sums, $\mu\mathcal{I}_A(M_1 + M_2) = \mathcal{I}_A(M_1)\mathcal{I}_A(M_2)$.*

PROOF. If M comes from S^3 by surgery on the framed link L, then $\overline{M}$ comes by surgery on $\overline{L}$, the image of L under an orientation reversing homeomorphism of S^3. Now $\overline{A} = A^{-1}$ and reflecting the crossing in Figure 2 exchanges the roles of A and A^{-1}. Recall, too, that Δ_n is real and that $S_n(\alpha)$ is a polynomial with integer coefficients. Then the first result follows. The second result comes from the fact that the connected sum of two manifolds comes by surgery from the separated union of the framed links that give the individual manifolds. □

In the above, a closed connected 3-manifold M was obtained from S^3 by surgery on an n-component framed link L with linking matrix Λ. If M itself contains a framed link K, this K can be regarded as being in $S^3 - L$ and isotopies of K in M correspond to isotopies of K in $S^3 - L$ together with 'Kirby-movements' of K across components of L. The embeddings of $S^1 \times D^2$ onto neighbourhoods of the components of L now induce a multilinear map

$$\langle\,,\,,\ldots,\,,K\rangle_L : \mathcal{S}(S^1 \times D^2) \times \mathcal{S}(S^1 \times D^2) \times \cdots \times \mathcal{S}(S^1 \times D^2) \to \mathcal{S}S^3$$

by evaluating in $\mathcal{S}S^3$ the images of n skeins from $S^1 \times D^2$ together with K.

THEOREM 4.2. *Let A be a primitive* $4r$*-th root of unity. An invariant of a framed link K in a closed orientable* 3*-manifold M is given by*

$$\mathcal{I}_A(M, K) = \mu\langle\mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r, K\rangle_L \langle\mu\omega_r\rangle_{U_-}^{\sigma A}.$$

This induces a well-defined linear map

$$\mathcal{I}_A(M, \ ): \mathcal{S}(M) \to \mathbb{C}.$$

5. Some applications

General skein theory considerations give immediately the following result.

LEMMA 5.1. *The map of framed links given by an embedding* $e: M_1 \to M_2$ *induces a map* $e: \mathcal{S}(M_1) \to \mathcal{S}(M_2)$ *which is linear if e is orientation preserving and semilinear if e is orientation reversing. Further, there is a natural identification*

$$\mathcal{S}(M_1 \sqcup M_2) = \mathcal{S}(M_1) \otimes \mathcal{S}(M_2).$$

The embedding of a disjoint union of solid tori onto the neighbourhood of a framed link has already illustrated both parts of this lemma. A second example is that, if F is any oriented surface, the embedding $(F \times I) \sqcup (F \times I) \to (F \times I)$, induced by including the two copies of the interval I onto the first and second halves of itself, induces an algebra structure on $\mathcal{S}(F \times I)$.

As a more extended example of this (and of Theorem 4.2) consider a handlebody H of genus g. The boundary of $H \times [-1, 1]$ is the connected sum of g copies of $S^1 \times S^2$ and so has invariant $\mu^{-(g-1)}$. There are natural disjoint embeddings of H onto $H \times \{-1\}$ and onto $H \times \{1\}$. One, the first say, is orientation preserving the other orientation reversing. By the previous remarks, these two embeddings induce a map

$$\mathcal{S}(H) \times \mathcal{S}(H) \to \mathcal{S}\big(\partial\big(H \times [-1, 1]\big)\big)$$

which is a sesquilinear map, meaning that, as a function of two variables, it is linear in the first variable and semilinear (that is, conjugate linear) in the second variable. For A a primitive $4r$-th root of unity, this map composed with $\mathcal{I}_A(\partial(H \times [-1, 1]), \ )$, becomes conjugate symmetric and so gives a Hermitian form on $\mathcal{S}(H)$. Let this form be denoted by $\phi: \mathcal{S}(H) \times \mathcal{S}(H) \to \mathbb{C}$. This defines a map from $\mathcal{S}(H)$ to its dual space. Let $\mathcal{S}(H)/\ker$ be the quotient of $\mathcal{S}(H)$ by the kernel of this map. In fact, as will be indicated later, $\mathcal{S}(H)/\ker$ is a finite-dimensional vector space. Further, at least when $A = \exp(\pi i/2r)$, it can be shown that ϕ is positive definite on $\mathcal{S}(H)/\ker$ and is hence a complex inner product on this quotient. A method for proving that will be outlined below, but for full details of this and related inner products see [48]. The boundary of a manifold always has a collar neighbourhood. Thus there is an embedding $H \sqcup (\partial H \times I) \to H$ which induces an action of $\mathcal{S}(\partial H \times I)$ on $\mathcal{S}(H)/\ker$ as an algebra of endomorphisms. Suppose now that $e_{\pm}: S^1 \times D^2 \to \partial H \times I$ are two embeddings onto a neighbourhood of a simple closed curve C in $\partial H \times \frac{1}{2}$ that send the natural framing of $S^1 \times D^2$ to (± 1)-framings of the neighbourhood of C (the zero framing being the natural product framing). Then $e_{\pm}(\mu\omega_r)$ are endomorphisms of $\mathcal{S}(H)/\text{ker}$ which will be denoted $\rho_C^{\pm}: \mathcal{S}(H)/\text{ker} \to \mathcal{S}(H)/\text{ker}$. Note that ρ_C^{+} and ρ_C^{-} are mutually inverse endomorphisms; this follows from the identity of Figure 13 and the definition of μ. If x is any element of $\mathcal{S}(H)$, consider the image under the sesquilinear map of $(\rho_C^{+}x, \rho_C^{+}x)$ in $\mathcal{S}(\partial(H \times [-1, 1]))$. For this two copies of $e_{+}: S^1 \times D^2 \to \partial H \times I \to H$ have been used followed by inclusions into $\partial(H \times [-1, 1])$. However the second inclusion is orientation *reversing*. Thus the second embedding can be isotoped out of $H \times \{1\}$ through $\partial H \times [-1, 1]$ into $H \times \{-1\}$ where it must be regarded as an instance of e_{-}. Then the combined effect of the two copies of $\mu\omega_r$ in $\mathcal{S}(\partial(H \times [-1, 1]))$ is that of $\rho_C^{+}\rho_C^{-}$ which is the identity map. Thus ρ_C^{+} is a unitary map with respect to the Hermitian form ϕ, namely $\phi(\rho_C^{+}x, \rho_C^{+}x) = \phi(x, x)$. Of course, any product of endomorphisms of the form $\rho_C^{\pm}$ is also unitary, because unitary automorphisms form a subgroup of the automorphisms of $\mathcal{S}(H)$. Suppose ρ is such a unitary automorphism. The Cauchy–Schwarz inequality implies for any x that $|\phi(x, \rho x)|^2 \leqslant \phi(x, x)\phi(\rho x, \rho x)$ and so $|\phi(x, \rho x)| \leqslant |\phi(x, x)|$.

THEOREM 5.1 (Originally in [7], see also [17] and [40]). *Suppose M is a genus g closed orientable 3-manifold. If $A = \exp(\pi i/2r)$ then*

$$|\mathcal{I}_A(M)| \leqslant \left|\sqrt{\frac{2}{r}} \sin \frac{\pi}{r}\right|^{1-g}.$$

PROOF. M can be obtained by gluing together two copies of a genus g handlebody H; so can the connected sum of g copies of $S^1 \times S^2$. The gluing map that gives this latter connected sum can be changed by a product of twist homeomorphisms of ∂H so that the new gluing gives M (see [21]). Thus M can be obtained from $\partial(H \times [-1, 1])$ by ± 1 surgery on simple closed curves each in some $\partial H \times t$. Thus a surgery link L, in S^3, for M is a surgery link for $\partial(H \times [-1, 1])$ (for example, the 0-framed unlink of g components) together with these extra surgery curves. Now $|\mathcal{I}_A(M)| = |\mu\langle \mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r\rangle_L|$ for $\langle \mu\omega_r\rangle_{U_-}$ has unit modulus. The right-hand side of this is, in the above notation, $|\phi(\emptyset, \rho\emptyset)|$ where $\emptyset$ denotes the empty skein in H and ρ is the composition of the $\rho_C^{\pm}$ where C runs through the above mentioned curves in the $\partial H \times t$. However $|\phi(\emptyset, \emptyset)|$ is just $|\mathcal{I}_A(\partial(H \times [-1, 1]))| = \mu^{-(g-1)}$. Hence the above instance of the Cauchy–Schwarz inequality implies that $|\mathcal{I}_A(M)| \leqslant |\mu^{-(g-1)}|$. But $A = \exp(\pi i/2r)$ implies that $\mu = \sqrt{\frac{2}{r}} \sin \frac{\pi}{r}$ and the result follows. □

This result, which gives a lower bound for the Heegaard genus of a 3-manifold, is at least one place where the quantum invariants interact with the concerns of classical 3-manifold theory. Y. Yokota develops these ideas in [48] where he considers skein spaces of a handlebody with some specified framed points on its boundary. The same general ideas concerning Hermitian forms lead him to the following generalisation of the above.

THEOREM 5.2. *Suppose a framed knot K in a closed orientable 3-manifold M meets each half of a genus g Heegaard splitting of M in n standard (boundary parallel) arcs. Let $A = \exp(\pi i/2r)$ then*

$$\left|\mathcal{I}_A\big(M, K(S_k(\alpha))\big)\right| \leqslant \left|\sqrt{\frac{2}{r}}\sin\frac{\pi}{r}\right|^{1-g}\left|\frac{\sin\frac{(k+1)\pi}{r}}{\sin\frac{\pi}{r}}\right|^n.$$

Here $\mathcal{I}_A\big(M, K(S_k(\alpha))\big)$ is the invariant of M including in it the skein $S_k(\alpha)$ about the knot K. A sharper inequality is available, [48], when M is a homology 3-sphere. Similarly these ideas can be interpreted in terms of the tunnel number of a knot in a 3-manifold (that is the minimal number of arcs that need to be added to the knot to produce a spine of one side of a Heegaard splitting) in the following way (see [48] and [17]).

THEOREM 5.3. *If K is a knot with tunnel number N in a closed oriented 3-manifold M and $A = \exp(\pi i/2r)$ then*

$$\sum_{k=0}^{r-2}\left|\mathcal{I}_A\big(M, K(S_k(\alpha))\big)\right|^2 \leqslant \left|\sqrt{\frac{2}{r}}\sin\frac{\pi}{r}\right|^{-2N}.$$

Again, a sharper result is available for a homology 3-sphere; the right-hand side can be multiplied by 2^{-N}.

The techniques just described were adapted in [35] to prove, for the first time, that certain Montesinos knots have tunnel number two. Suppose α is an arc added to a knot K in S^3, so that $\alpha \cup K$ is a spine of a handlebody H that is one side of a Heegaard splitting of S^3. Then there is an involution ι of S^3, leaving K invariant and having fixed point set an unknot U that meets K in two points and H in three intervals. Then S^3/ι is also a copy of S^3 with H/ι a ball. $(K \cup U)/\iota$ is a, probably knotted, θ-curve meeting $\partial H/\iota$ in six points. It is the skein theory of the ball with six sets of boundary points (each labelled with an idempotent of a Temperley–Lieb algebra) that is considered. Pairing this with the similar space for the complementary ball gives a Hermitian form and (with some care with the labellings) elements of the six string braid group give unitary automorphisms. The resulting Cauchy–Schwarz inequality concerns the evaluation in $\mathcal{S}S^3$ of the above θ-curve with labelled edges and triads at the vertices. However, symmetries of Montesinos knots are classified so this can be checked against evaluations on the only possible θ-curves that can arise. A contradiction sometimes results. Although this proof uses all the technology of triads and idempotents, it does not really use the 3-manifold invariants, for all the calculations are in S^3 and ω_r is never mentioned.

Towards the end of [14], Kirby and Melvin study these $\mathrm{SU}_q(2)$-invariants, when A is a primitive $4r$-th root of unity, for small values of r. They make use of the fact that the Jones polynomial at $t = e^{2\pi i/r}$ is, for certain small values of r, expressible in terms of more classical invariants [30]. They also employ a Symmetry Principle (see also [25]) which has the effect of reducing the number of terms in the expansion of ω_r. When $r = 3$, $\mathcal{I}_A(M)$ is expressed in terms of the Brown invariant of the linking matrix of a surgery link and of the dimension of $H^1(M; \mathbb{Z}/2\mathbb{Z})$. When $r = 4$, $\mathcal{I}_A(M)$ is simply expressed in terms of the

μ-invariants of all the spin structures on M. When $r = 5$ there is, when M is a homology 3-sphere gained by Dehn surgery on a knot, a correlation between $\mathcal{I}_A(M)$ and the Casson invariant of M modulo 5. The following result, also, appears in [14].

THEOREM 5.4. *Suppose the Jones polynomials of knots K_1 and K_2 have distinct evaluations at $e^{2\pi i/5}$. Then, for every integer n, surgery on S^3 along K_1 and K_2, each with framing n, gives distinct oriented 3-manifolds.*

In the preamble to Theorem 5.1, with A a primitive $4r$-th root of unity, it was shown how a twist homeomorphism about a simple closed curve C in the boundary of a handlebody H led to automorphisms $\rho_C^{\pm}$ of $\mathcal{S}(H)/\text{ker}$. This induces a projective representation of the mapping class group of H on $\mathcal{S}(H)/\text{ker}$ (see [40]). It is only projective as powers of the unit modulus complex number $\langle \mu\omega_r \rangle_{U_-}$ have to be 'overlooked'. That however is rectified in [33] where a genuine representation, of an extension by $\mathbb{Z}$ of the mapping class group, is obtained essentially in this way. In [16] 3-manifold invariants obtained by means of representations of mapping class groups are discussed. Further, it is shown in [5] that $\mathcal{S}(H)/\text{ker}$ is the vector space that should be associated with the surface ∂H in a Topological Quantum Field Theory that is associated with the theory of this $\text{SU}_q(2)$ invariant of closed 3-manifolds. Roughly, the idea is that a cobordism between ∂H_1 and ∂H_2 can be obtained by surgery on a framed link in H_2 that lies in the complement of a copy of H_1. Placing ω_r about each surgery curve and an arbitrary skein in the copy of H_1 should lead to an element of $\mathcal{S}(H_2)$. In this way the cobordism might lead to a linear map $\mathcal{S}(H_1)/\text{ker} \to \mathcal{S}(H_2)/\text{ker}$, that being the essence of a Topological Quantum Field Theory. There are however, difficulties with the term in $\langle \mu\omega_r \rangle_{U_-}$ that occurs in the definition of $\mathcal{I}_A(M)$ that necessitate the supposition that each cobordism be equipped with a Pontryagin class (or 'given a p_1-structure'). The details are fully explained in [5].

6. Fusion

To proceed further one must use a little more daring with the Temperley–Lieb algebras and their idempotents than has so far been attempted here. Although some progress can be made without any quotients, [27], it will usually be appropriate to work with skeins modulo $f^{(r-1)}$.

The first diagram of Figure 14 shows a 'triad' element of the skein space of a ball B with $a + b + c$ framed points specified on its boundary. The diagram will be abbreviated, to a black dot with three labelled protruding edges, as shown. Note the presence of the idempotents of the various TL_n and note that if $a = 0$ then $b = c$ and the triad is just $f^{(b)}$. If $B \subset S^3$ then the closure of $S^3 - B$ is also a ball and a base of the skein of this outside ball is (projecting onto a disc) represented by diagrams of arcs with no crossing. Combining such a base element with the triad will always give the zero element of $\mathcal{S}S^3$ except possibly when a similar triad is placed in the outside ball and there results the element $\theta(a, b, c)$ as shown.

A little work [28,34] and [11], shows that

$$\theta(a,b,c) = (\Delta_{x+y+z}!\Delta_{x-1}!\Delta_{y-1}!\Delta_{z-1}!)/(\Delta_{y+z-1}!\Delta_{z+x-1}!\Delta_{x+y-1}!),$$

Fig. 14.

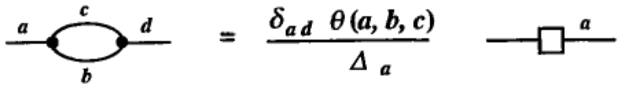


Fig. 15.

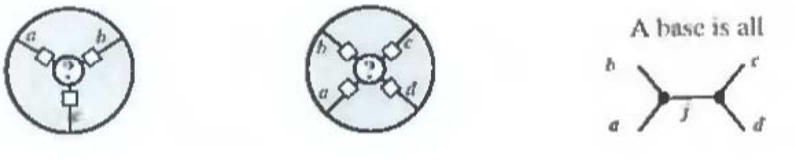


Fig. 16.

where $\Delta_n!$ denotes $\Delta_n\Delta_{n-1}\Delta_{n-2}\ldots\Delta_1$, this being 1 if n is -1 or zero. It follows, when A is a primitive $4r$-th root of unity, that it is precisely when (a, b, c) is r-admissible in the following sense that the triad exists and $\theta(a, b, c) \neq 0$.

DEFINITION. The triple of non-negative integers (a, b, c) is r-admissible provided that $a + b + c$ is even, $a \leqslant b + c$, $b \leqslant c + a$, $c \leqslant a + b$ and $a + b + c \leqslant 2(r - 2)$.

All black dots in the following diagrams will refer to r-admissible triples, and the phrase 'modulo r' will mean 'modulo r-inadmissible triples' (which include $f^{(r-1)}$).

The identity of Figure 15 is an easy exercise.

The left-hand diagram of Figure 16 symbolises the subspace of the skein space of the ball, with $a + b + c$ framed boundary points, that has idempotents as shown and any skein element in the middle smaller ball. That subspace has dimension one or zero, according as to whether an (a, b, c) triad exists. Modulo r it has dimension one if and only if (a, b, c) is r-admissible. The right-hand diagram similarly refers to the ball with $a+b+c+d$ framed boundary points. The subspace here has, modulo r, a base of all r-admissible elements as shown. To show these elements span employ the fact that, in TL_{r-1}, $1 = f^{(r-1)} + E$ where E is a sum of products of the e_i. Here $f^{(r-1)}$ can be forgotten, so a link with $r - 1$ parallel arcs travelling from right to left can be replaced with a sum of links with fewer arcs going all the way from right to left. Independence follows by adjoining a triad to a postulated linear dependence. Details are in [28,34] and [11].

This base has a 'horizontal bias' so there is another base with a 'vertical bias'. The change of base formula is in Figure 17, the summation being over all i for which both the triads in the diagram are r-admissible. The terms of the change of base matrix $\{\begin{smallmatrix} a & b & i \\ c & d & j \end{smallmatrix}\}$ are

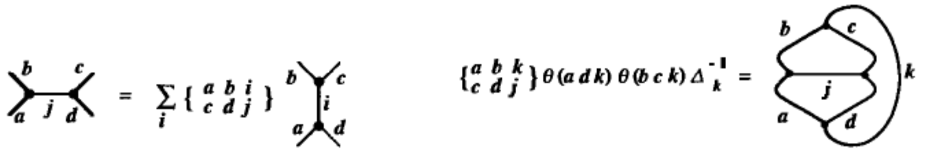


Fig. 17.

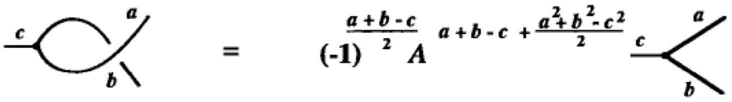


Fig. 18.

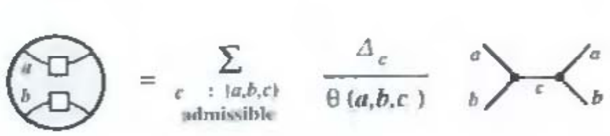


Fig. 19.

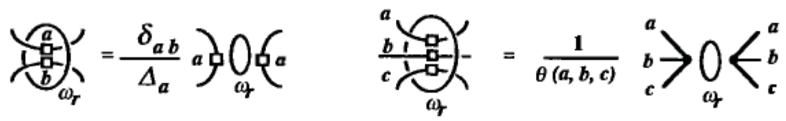


Fig. 20.

traditionally called '$6j$-symbols'. Adjoining triads to the change of base formula easily produces the expression shown for a $6j$-symbol, in terms of a framed, labelled, tetrahedral graph with triads at the vertices, regarded as an element of $\mathcal{S}S^3 \equiv \mathbb{C}$.

In theory one knows these $6j$-symbols. A lengthy, closed formula is known for them [3, 11,15]. The equality of Figure 18 is, however, a routine exercise.

Some expression of the form of Figure 19 must be true. The coefficients come easily by adjoining triads.

In principle, the formulae of Figures 15, 17–19, together, of course, with knowledge of formulae for Δ_n, $\theta(a, b, c)$ and (regrettably) the $6j$-symbols, give all that is required to calculate, modulo $f^{(r-1)}$, the value in $\mathcal{S}S^3$ of any framed link with components decorated with various $S_n(\alpha)$'s. Then $\mathcal{I}_A(M)$ is a linear sum of such terms. Use Figure 19 to fuse strands near crossings, Figure 18 to remove the crossings, Figure 17 to change trivalent labelled framed graphs until they contain 2-gons, then remove 2-gons using Figure 15. There results an awesome sum of products of the Δ_n, $\theta(a, b, c)$ and $6j$-symbols.

The formulae of Figure 20, when A is a primitive $4r$-th root of unity, follow easily from Figures 19 and 12. The right-hand expression is to be taken to be zero unless (a, b, c) is r-admissible.

Fig. 21.

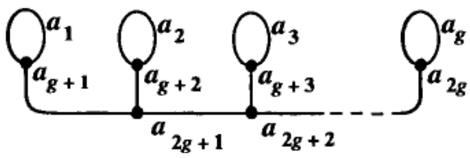


Fig. 22.

Suppose one applies Figure 19 and then Figure 15 to two zero framed curves encircling $S^1 \times D^2$ labelled with $S_a(\alpha)$ and $S_b(\alpha)$. Then one obtains the following (Clebsch–Gordan) formula in $\mathcal{S}(S^1 \times D^2)$ modulo r.

$$S_a(\alpha)S_b(\alpha) = \sum_c S_c(\alpha)$$

the sum being over all c such that (a, b, c) is r-admissible.

Application of Figure 20 to the element of $\mathcal{S}(S^1 \times D^2)$ shown in Figure 21 shows that element to be $\mu^{-2}\sum_{a=0}^{r-2}(S_a(\alpha))^2$. Suppose g copies of Figure 21 are threaded (like beads) on a 0-framed ω_r-decorated unknot in S^3. The resulting element of $\mathcal{S}(S^3)$, modulo r, is, using the above Clebsch–Gordan formula and Figure 12, $\mu^{-2(g+1)}N_g$ where N_g is the number of ways that the Clebsch–Gordan formula produces an $S_0(\alpha)$ term from $(\sum_{a=0}^{r-2}(S_a(\alpha))^2)^g$.

This N_g is just the number of ways of labelling the edges of the graph of Figure 22 with non-negative integers $\{a_i\}$ so that an r-admissible triad appears at each vertex.

THEOREM 6.1. *Let F be a closed orientable surface of genus g and let A be a primitive $4r$-th root of unity. Then $\mathcal{I}_A(S^1 \times F) = N_g$.*

PROOF. A surgery diagram for $S^1 \times F$ is the sum of g copies of the 0-framed Borromean rings (with one component featuring in every summation). It has just been shown that, in $\mathcal{S}S^3$, this framed link, with every component decorated with ω_r, is $\mu^{-2(g+1)}N_g$. To obtain $\mathcal{I}_A(S^1 \times F)$ this must be multiplied by a μ for each link component and one more for convention. (Note that the framed link has zero linking matrix which has zero signature.) The result follows. □

This result means, for $g = 1$, that $\mathcal{I}_A(S^1 \times S^1 \times S^1) = (r - 1)$. A similar calculation applied to the link of Figure 23 shows that the manifold gained from it by surgery has the

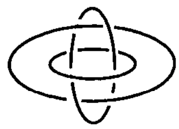


Fig. 23.

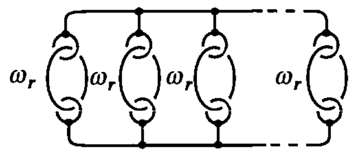


Fig. 24.

same invariant. That manifold is the $(S^1 \times S^1)$-bundle over S^1 that has minus the identity as monodromy map. The first homology group distinguishes this from $S^1 \times S^1 \times S^1$.

Now regard the graph of Figure 22 as the spine of a genus g handlebody H. Each labelling of the edges of the graph, with r-admissible triples at vertices, gives, on inserting triads at the vertices and connecting them all up with parallel arcs along the edges, an element of $\mathcal{S}H$. The set of elements coming from all such labellings spans $\mathcal{S}H$ modulo r. This follows by a similar argument to that outlined above for the ball with $a+b+c+d$ points in its boundary (see Figure 16). There are two copies of this graph in Figure 24.

Inserting skeins into the neighbourhoods of the two graphs and evaluating in $\mathcal{S}S^3$ gives a pairing $\mathcal{S}H \times \mathcal{S}H \to \mathbb{C}$. When A is a primitive $4r$-th root of unity, this pairing is an orthogonal pairing on the set of skeins given by r-admissible labellings of the graphs. That can, at once, be used to show that skeins coming from distinct such labellings are independent and hence form a base (more details are in [28] and [34]). Thus N_g is the dimension of $\mathcal{S}H$ modulo r. If the handlebody neighbourhood of just one of the graphs is given an orientation disagreeing with that of S^3, this pairing is the same as the pairing

$$\phi : \mathcal{S}(H)/\ker \times \mathcal{S}(H)/\ker \to \mathbb{C}$$

already considered for Theorem 5.1. Now using the given base elements it is straightforward, using Figure 20, to make calculations (as required in Theorem 5.1) on Figure 25 to determine when ϕ is positive definite.

7. The state sum invariants of Turaev and Viro

Other invariants of a closed orientable 3-manifold M, closely allied to the invariants just discussed, are the Turaev–Viro invariants [44]. Many calculations of these invariants can be found in [11]. They were defined as state sums, as sums over all possible 'states' of the products of some 'weights' associated to the state. Invariance was checked either by using the Matveev moves on a special spine of the 3-manifold, or by using the Pachner moves [38] on a triangulation. Later Turaev proved that this invariant is the square of the modulus of a corresponding $SU_q(2)$-invariant. Though in one way that is disappointing, in that the Turaev–Viro invariant is therefore a weaker invariant, the greater calculability of a state sum might lead to applications. This has all been re-cast into a very simple form, by Roberts [41], in the following way.

Consider first a genus g Heegaard splitting diagram for M. This is taken to be two *disjoint* sets of curves in a handlebody H: a complete set in ∂H of meridian curves $\{\delta_i\}$ for H, and a complete set of meridian curves $\{\varepsilon_i\}$ for the closure of $M - H$ *pushed into H* along a collar of ∂H. All the curves inherit framings from ∂H. The framed link $L(s)$ in S^3, corresponding to this splitting diagram, is that consisting of the images of these curves under a standard embedding of H into S^3. (Actually, any embedding will do, see [41].) The signature of the linking matrix of $L(s)$ is zero as the matrix contains a $g \times g$ block of zeros. Further, $L(s)$ is a surgery diagram for $M + \overline{M}$. This can be seen by the Kirby-calculus technique of regarding the surgery process as giving a 3-manifold as the the boundary of a 4-ball with added handles; regard the δ-curves as giving 1-handles and the ε-curves as being attaching circles for 2-handles [13]. It follows from Lemma 4.1 that

$$\left|\mathcal{I}_A(M)\right|^2 = \mu^2\langle \mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r\rangle_{L(s)}.$$

Now consider the 3-manifold M to be triangulated with d_i i-simplexes. A regular neighbourhood of the *dual* 1-skeleton is a handlebody H. Take on ∂H a curve δ_i that is the intersection of ∂H with the i-th 2-simplex. Let ε_i be the result of pushing a little into H a curve in ∂H that bounds a disc in $\overline{M - H}$ which intersects transversely the 1-skeleton at the barycentre of the i-th 1-simplex. Framings on the curves come from ∂H. Now regard H as embedded in some standard way in S^3 and let $L(t)$ be the framed link in S^3 of all the ε and δ curves coming from the triangulation in the above way. This is not quite the link from a Heegaard splitting; some of its curves must be deleted, namely the $d_3 - 1$ δ-curves corresponding to a maximal tree of the dual 1-skeleton and the $d_0 - 1$ ε-curves corresponding to a maximal tree of the 1-skeleton. Let $L(t)^-$ be the link after deletions. If all components of $L(t)$ are decorated with $\mu\omega_r$ those corresponding to curves not in $L(t)^-$ can be slid over curves of $L(t)^-$, into small balls, without changing the resulting element of $\mathcal{S}S^3$. Thus

$$\left|\mathcal{I}_A(M)\right|^2 = \mu^{d_0+d_3}\langle \mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r\rangle_{L(t)}.$$

The link $L(t)$ is Roberts' 'chain mail' link, the reference being to mediaeval armour. Certainly, if one constructs this link from a fine subdivision of a given triangulation it is a link that occupies (protects) the manifold in a ubiquitous manner; operations on it might be reminiscent of integration.

Notice that in the link $L(t)$ a δ-curve bounds a disc in H (namely H intersected with a 2-simplex) that meets three and only three ε-curves. To evaluate $\langle \mu\omega_r, \mu\omega_r, \ldots, \mu\omega_r\rangle_{L(t)}$ consider the multilinear expansion that occurs on substituting $\omega_r = \sum_{n=0}^{r-2} \Delta_n S_n(\alpha)$ for each ε-curve but leaving ω_r 'intact' on the δ-curves. Then use the right-hand identity of Figure 20 on each labelled diagram of the resulting summation (recall that, in Figure 20, each side of the identity is non-zero only when (a, b, c) is r-admissible). Each labelled diagram then breaks up into many labelled tetrahedral graphs and 0-framed unknots labelled ω_r. Tidying up the details, Roberts obtains the following theorem. The state summation expression here given for $|\mathcal{I}_A(M)|^2$ is, up to normalisation conventions, the Turaev–Viro invariant of M.

THEOREM 7.1. *Let M be a closed orientable 3-manifold with a triangulation having d_0 0-simplexes. A state is a function s from the set of edges of the triangulation to $\{0, 1, 2, \ldots, r-2\}$ that maps the three edges of any 2-simplex to an r-admissible triple. Then, if A is a primitive $4r$-th root of unity,*

$$|\mathcal{I}_A(M)|^2 = \mu^{2d_0} \sum_s \left\{ \prod_e \Delta_{se} \prod_f \theta(sf) \prod_t \tau(st) \right\}.$$

The summation is over all states, e, f and t are the edges, faces and tetrahedra of the triangulation, s maps the edges of face f to the triple sf. Similarly $\tau(st)$ is the skein evaluation of a tetrahedral graph, with triads at the vertices and edges labelled with the values of s on the edges of the tetrahedron t (see Figure 17).

8. Invariants from HOMFLY-based skeins

Quantum 3-manifold invariants corresponding to SU(N) for $N > 2$ have been substantiated by Turaev and Wenzl [45] using intricacies of representation theory. More recently a more geometric skein theory proof of the existence of these invariants has been developed by Yokota [49]. The proof is self-contained though clearly inspired by representation theory. An outline of it now follows. Throughout, N will be a fixed integer, $N \geqslant 2$ and $t \in \mathbb{C}$ a primitive $2N(K+N)$ root of unity for some 'level' K. Let M be an oriented 3-manifold together with a finite set of framed points in ∂M, each such point being oriented (that is, it has assigned to it a direction 'into' or 'out of' the manifold).

DEFINITION. The linear skein $\mathcal{S}_N(M)$ is the $\mathbb{C}$-vector space of formal linear sums of framed oriented links in M, that consist of closed curves and arcs that meet ∂M in precisely the given framed oriented points, quotiented by: (i) isotopy of framed links; (ii) $L \cup \bigcirc = [N]L$, where $[k] = (t^{kN} - t^{-kN})/(t^N - t^{-N})$; (iii) adding ± 1 to the framing of any component of L changes L to $t^{\pm(N^2-1)}L$; (iv) the skein identity of Figure 25.

The familiar theory of the HOMFLY polynomial, [31] for example, asserts that again the skein space of S^3 is one-dimensional and that, if the 0-framed unknot is taken as its

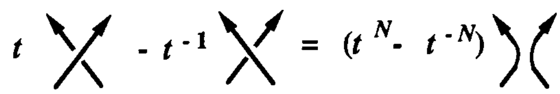


Fig. 25.

base, the coordinate of any 0-framed knot L is the HOMFLY polynomial of L with the two variables taken to be $\mathrm{i}t^{N^2}$ and $\mathrm{i}(t^{-N} - t^N)$, in the notation of [31]. This is well known to be the substitution corresponding to $\mathrm{SU}(N)$. Nevertheless, in what follows *the empty diagram* is taken as base for $\mathcal{S}_N(S^3)$ and is used to identify it with $\mathbb{C}$.

The skein space of the cube with l in-going points on its bottom face and l out-going points on its top face will be $\mathcal{S}_N(B_l^l)$. If $\xi, \eta \in \mathcal{S}_N(B_l^l)$, placing ξ above η induces a well defined product and $\mathcal{S}_N(B_l^l)$ becomes an (Hecke) algebra. If $\xi \in \mathcal{S}_N(B_k^k)$ and $\eta \in \mathcal{S}_N B_l^l$ then juxtaposing cubes side by side produces $\xi \otimes \eta \in \mathcal{S}_N(B_{k+l}^{k+l})$. Similarly a tensor product in $\mathcal{S}_N(S^1 \times D^2)$ comes by embedding two solid tori in one. If $\xi \in \mathcal{S}_N(B_l^l)$ then $\widehat{\xi}$ denotes the element of $\mathcal{S}_N(S^1 \times D^2)$ obtained by placing the cube containing ξ in the solid torus in a standard way and joining the l points at the top of the cube to those at the bottom of the cube by l embedded arcs each encircling the solid torus in a positive direction. The association $\xi \mapsto \widehat{\xi}$ induces a linear map $\mathcal{S}_N(B_l^l) \to \mathcal{S}_N(S^1 \times D^2)$. Another linear map $\chi : \mathcal{S}_N(S^1 \times D^2) \to \mathbb{C}$ comes from taking an embedding, preserving orientation and framing, of $S^1 \times D^2$ onto a neighbourhood of a 0-framed unknot.

Suppose that σ_i is the usual i-th generator of the l-string braid group which provides diagrams for elements of $\mathcal{S}_N(B_l^l)$. Regarding σ_i as an element of $\mathcal{S}_N(B_l^l)$, it is clear that $\{\sigma_i^{\pm 1}\colon 1 \leqslant i \leqslant l-1\}$ is a set of generators for $\mathcal{S}_N(B_l^l)$ as an algebra. Yokota [49] focusses attention on two special elements of $\mathcal{S}_N(B_l^l)$. Both are idempotents. The first $f^{(l)}$, the symmetriser, has the property that for each i,

$$\sigma_i^{\pm 1} f^{(l)} = t^{\pm(N-1)} f^{(l)} = f^{(l)} \sigma_i^{\pm 1},$$

the anti-symmetriser $g^{(l)}$ is such that

$$\sigma_i^{\pm 1} g^{(l)} = -t^{\mp(N+1)} g^{(l)} = g^{(l)} \sigma_i^{\pm 1}.$$

These are combined together to form further idempotents corresponding to Young diagrams. A Young diagram λ is a way of partitioning a finite set of $|\lambda|$ elements, represented by little squares, in two related ways; the sets of the partitions are the rows and columns of a diagram of little squares as in Figure 26.

The Young diagram λ is specified by a sequence $[l_{N-1}, l_{N-2}, \dots, l_2, l_1]$ of non-negative integers where there are $l_{N-1} + l_{N-2} + \cdots + l_i$ squares in the i-th row (the top row being the first); there are $N-1$ rows though the final rows may be empty. The idempotent $e_\lambda \in \mathcal{S}_N(B_{|\lambda|}^{|\lambda|})$ is a scalar multiple of the element formed in the following way. Take an anti-

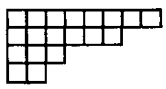


Fig. 26.

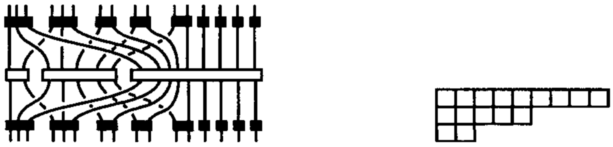


Fig. 27.

symmetriser of the size of each column of the diagram and place them side by side to form

$$\left(g^{(N-1)}\right)^{\otimes l_{N-1}} \otimes \left(g^{(N-2)}\right)^{\otimes l_{N-2}} \otimes \cdots \otimes \left(g^{(1)}\right)^{\otimes l_1},$$

connect this to the element

$$f^{(l_{N-1})} \otimes f^{(l_{N-1}+l_{N-2})} \otimes \cdots \otimes f^{(l_{N-1}+l_{N-2}+\cdots+l_1)}$$

formed by juxtaposing symmetrisers with sizes the rows of the diagram, and connect that to a second tensor product of anti-symmetrisers, identical to the one first mentioned, in the way shown in Figure 27 (where the g-idempotents are black rectangles and the f-idempotents are white ones).

The complex number Δ_λ is defined by

$$\Delta_\lambda = \chi(\widehat{e_\lambda}).$$

In [49] it is, at some inconvenience, shown that

$$\Delta_\lambda = \prod_{n=1}^{N-1} \prod_{i=1}^{n} [l_n + l_{n-1} + \cdots + l_i + n - i + 1]/[n - i + 1].$$

Now, for K a fixed positive integer the set of Young diagrams $\Gamma_{N,K}$ is defined by

$$\Gamma_{N,K} = \left\{ \lambda\colon \sum_{i=1}^{N-1} l_i \leqslant K \right\}$$

and the element $\Omega_K \in \mathcal{S}_N(S^1 \times D^2)$ is defined by

$$\Omega_K = \sum_{\lambda \in \Gamma_{N,K}} \Delta_\lambda \widehat{e_\lambda}.$$

Recall that t is a primitive $2N(N+K)$ root of unity. The main result of Yokota [49] is that, if Ω_K occurs as the decoration on any component C of a framed link in S^3, then that link may be changed by pushing (according to the fashion of a Kirby K(2) move) any other component with either orientation across C, and (retaining the Ω_K decoration on C) the evaluation of the decorated link in $\mathcal{S}_N(S^3)$ is not altered. Thus in this situation Ω_K has exactly the same property as had ω_r in the previous discussion. Then, in the same way as before, an $SU_q(N)$ invariant for an oriented 3-manifold M is constructed as is stated in the next theorem (from [49]).

THEOREM 8.1. *Suppose the 3-manifold M is obtained by surgery on a framed link L in S^3. A well-defined invariant $\mathcal{I}_{N,K}(M)$ is defined by*

$$\mathcal{I}_{N,K}(M) = \theta \langle \theta\Omega_K \rangle^{\sigma}_{U_-} \langle \theta\Omega_K, \theta\Omega_K, \ldots, \theta\Omega_K \rangle_L.$$

Here $\langle\ ,\ ,\ldots,\ \rangle_L$ is the multilinear map on $\mathcal{S}_N(S^1 \times D^2)$ coming from L, U_- is the unknot with framing -1, the signature of the linking matrix of L is σ and θ is a normalising factor defined so that $\theta^{-2} = \langle \Omega_K \rangle_U$ where U is the 0-framed unknot. As before

$$\langle \Omega_K \rangle_{U_+} \langle \Omega_K \rangle_{U_-} = \langle \Omega_K \rangle_U = \sum_{\lambda \in \Gamma_{N,K}} \Delta_\lambda^2$$

and that is non-zero.

Yokota observes that if Ω_K^* is defined to be the same as Ω_K but with every string orientation in its construction reversed, then Ω_K^* may be used in place of any (or every) occurrence of Ω_K, without changing the outcome in any way. That is because Ω_K^* will still have the fundamental property that, in any decorated framed link in S^3, strings maybe slid over Ω_K^* without changing evaluations in $\mathcal{S}_N(S^3)$. So in $\mathcal{S}_N(S^1 \times D^2)$, modulo the effect of the root of unity, $\Omega_K^* \Omega_K = \langle \Omega_K \rangle_U \Omega_K^* = \langle \Omega_K^* \rangle_U \Omega_K$ by performing each of the permitted slidings. However changing *all* the orientations of a link never changes its HOMFLY polynomial. Thus $\langle \Omega_K \rangle_U = \langle \Omega_K^* \rangle_U$ and, modulo the root of unity, $\Omega_K^* = \Omega_K$.

Clearly, investigation of the $SU_q(N)$ invariants can be attempted along the lines used above for the $SU_q(2)$ invariants. However the necessary coefficients are, at every point, distinctly more cumbersome. In a way similar to that of Theorem 6.1 it can be shown (see [29]) that

$$\mathcal{I}_{N,K}(S^1 \times S^1 \times S^1) = |\Gamma_{N,K}|.$$

Similarly, for the $(S^1 \times S^1)$-bundle over S^1 obtained by surgery on Figure 23,

$$\mathcal{I}_{N,K}(S^1 \tilde{\times} (S^1 \times S^1)) = \left|\{\lambda \in \Gamma_{N,K}\colon \lambda = \lambda^*\}\right|.$$

Here λ^* is the Young diagram dual to λ; if λ is specified by $[l_{N-1}, l_{N-2}, \ldots, l_1]$ then λ^* is specified by $[l_1, l_2, \ldots, l_{N-1}]$. When $N = 2$ every λ is self-dual but otherwise that is not so. Thus these manifolds are distinguished by their $SU_q(N)$ invariants for $N > 2$ but not by their $SU_q(2)$ invariants. For lens spaces calculations can be made in the same way as before but the results are not immediately attractive; some of the necessary formulae for doing this, obtained from a perspective different from skein theory, can be found in [18].

9. Conclusion

This article has attempted to give some details of some of the simplest aspects of the theory of quantum invariants. The result is but a scratching of the surface of a vibrant, far-reaching, rapidly developing subject that has gained a momentum of its own. Because this topic interacts with many disciplines it has been interpreted in many different ways with respect to varying agenda. Recently there has been considerable investigation into the behaviour of the $SU_q(2)$ invariants as r (in the above notation) tends to infinity; that is Witten's 'large k limit'. An interpretation by Ohtsuki [37] led to a formulation of 'finite type' 3-manifold invariants, which are an analogue of the Vassiliev invariants in knot theory [2]. A survey of that topic is in [32]. The analogue is followed further in [19] where a universal 3-manifold invariant is constructed. This gives, for a 3-manifold, a class of formal sums of trivalent graphs to which the combinatorics of a Lie algebra may be applied to obtain a numerical invariant; again, surgery and the Kirby calculus provide the contact with actual 3-manifolds and the universal Vassiliev–Kontsevich invariant for knots is the inspiration. In all of the quantum invariant theory it is remarkable that Witten's original formula has led to so much creativity and there is the feeling that there are more insights to come.

References

[1] M.F. Atiyah, *The Geometry and Physics of Knots*, Cambridge Univ. Press (1990).

[2] D. Bar-Natan, *On the Vassiliev knot invariants*, Topology **34** (1995), 423–472.

[3] C. Blanchet, *Invariants on three-manifolds with spin structure*, Comment. Math. Helv. **67** (1992), 406–427.

[4] C. Blanchet, N. Habegger, G. Masbaum and P. Vogel, *Three-manifold invariants derived from the Kauffman bracket*, Topology **31** (1992), 685–699.

[5] C. Blanchet, N. Habegger, G. Masbaum and P. Vogel, *Topological quantum field theories derived from the Kauffman bracket*, Topology **34** (1995), 883–927.

[6] D.S. Freed and R.E. Gompf, *Computer calculations of Witten's 3-manifold invariants*, Comm. Math. Phys. **141** (1991), 79–117.

[7] S. Garoufalidis, *PhD dissertation*, University of Chicago (1992).

[8] L.C. Jeffrey, *Chern–Simons–Witten invariants of lens spaces and torus bundles, and the semi-classical approximation*, Comm. Math. Phys. **147** (1992), 563–604.

[9] J. Kania-Bartoszynska, *Examples of different 3-manifolds with the same invariants of Witten and Reshetikhin*, Topology **32** (1993), 47–54.

[10] C. Kassel, *Quantum Groups*, Springer (1995).

[11] L.H. Kauffman and S. Lins, *Temperley–Lieb Recoupling Theory and Invariants of 3-Manifolds*, Ann. of Math. Stud., Vol. 134, Princeton University Press, Princeton, NJ (1994).

[12] R.C. Kirby, *A calculus for framed links in S^3*, Invent. Math. **45** (1978), 35–56.

[13] R.C. Kirby, *The Topology of 4-Manifolds*, Lecture Notes in Math., Vol. 1374, Springer (1989).

[14] R.C. Kirby and P. Melvin, *The 3-manifold invariants of Witten and Reshetikhin–Turaev for* $\mathrm{sl}(2, \mathbb{C})$, Invent. Math. **105** (1991), 473–545.

[15] A.N. Kirillov and N.Y. Reshetikhin, *Representations of the algebra $U_q(\mathrm{Sl}(2))$, q-orthogonal polynomials and invariants of links*, Infinite Dimensional Lie Algebras and Groups, Adv. Ser. Math. Phys., Vol. 7, World Scientific (1989), 285–339.

[16] T. Kohno, *Topological invariants of three-manifolds using representations of mapping class groups I*, Topology **31** (1992), 203–230.

[17] T. Kohno, *Tunnel numbers of knots and Jones–Witten invariants. Braid groups, knot theory and statistical mechanics II*, Adv. Ser. Math. Phys., Vol. 17, World Scientific (1994), 275–293.

[18] T. Kohno and T. Takata, *Symmetry of Witten's 3-manifold invariants for* sl$(n, \mathbb{C})$, J. Knot Theory Ramifications **2** (1993), 149–169.
[19] T.T.Q. Le, J. Murakami and T. Ohtsuki, *On a universal perturbative invariant of 3-manifolds*, Topology **37** (1998), 539–574.
[20] B.-H. Li and T.-J. Li, *Generalized Gaussian sums and Chern–Simons–Witten–Jones invariants of lens spaces*, J. Knot Theory Ramifications **5** (1996), 183–224.
[21] W.B.R. Lickorish, *A representation of orientable combinatorial 3-manifolds*, Ann. of Math. **76** (1962), 531–540.
[22] W.B.R. Lickorish, *Surgery on knots*, Proc. Amer. Math. Soc. **60** (1977), 296–298.
[23] W.B.R. Lickorish, *Invariants for 3-manifolds from the combinatorics of the Jones polynomial*, Pacific J. Math. **149** (1991), 337–347.
[24] W.B.R. Lickorish, *Three-manifolds and the Temperley–Lieb algebra*, Math. Ann. **290** (1991), 657–670.
[25] W.B.R. Lickorish, *Calculations with the Temperley–Lieb algebra*, Comment. Math. Helv. **67** (1992), 571–591.
[26] W.B.R. Lickorish, *Distinct 3-manifolds with all* $\mathrm{SU}(2)_q$ *invariants the same*, Proc. Amer. Math. Soc. **117** (1993), 285–292.
[27] W.B.R. Lickorish, *The skein method for three-manifold invariants*, J. Knot Theory Ramifications **2** (1993), 171–194.
[28] W.B.R. Lickorish, *Skeins and handlebodies*, Pacific J. Math. **159** (1993), 337–349.
[29] W.B.R. Lickorish, *Sampling the* $\mathrm{SU}(N)$ *invariants of three-manifolds*, J. Knot Theory Ramifications **6** (1997), 45–60.
[30] W.B.R. Lickorish and K.C. Millett, *Some evaluations of link polynomials*, Comment. Math. Helv. **61** (1986), 349–359.
[31] W.B.R. Lickorish and K.C. Millett, *A polynomial invariant of oriented links*, Topology **26** (1987), 107–141.
[32] X.-S. Lin, *Finite type invariants of integral homology 3-spheres: a survey*, Banach Centre Publ., Vol. 42, Polish Acad. Sci., Warsaw (1998), 205–220.
[33] G. Masbaum and J.D. Roberts, *On central extensions of mapping class groups*, Math. Ann. **302** (1995), 131–150.
[34] G. Masbaum and P. Vogel, *3-valent graphs and the Kauffman bracket*, Pacific J. Math. **164** (1994), 361–381.
[35] K. Morimoto, M. Sakuma and Y. Yokota, *Identifying tunnel number one knots*, J. Math. Soc. Japan **48** (1996), 667–688.
[36] J.R. Neil, *Combinatorial calculations of the various normalisations of the Witten invariants for 3-manifolds*, J. Knot Theory Ramifications **1** (1992), 407–499.
[37] T. Ohtsuki, *A polynomial invariant of integral homology 3-spheres*, Math. Proc. Cambridge Philos. Soc. **117** (1995).
[38] U. Pachner, *P.L. homeomorphic manifolds are equivalent by elementary shellings*, Europ. J. Combinatorics **12** (1991), 129–145.
[39] N.Y. Reshetikhin and V.G. Turaev, *Invariants of 3-manifolds via link polynomials and quantum groups*, Invent. Math. **103** (1991), 547–597.
[40] J.D. Roberts, *Skeins and mapping class groups*, Math. Proc. Cambridge Philos. Soc. **115** (1994), 53–77.
[41] J.D. Roberts, *Skein theory and Turaev–Viro invariants*, Topology **34** (1995), 771–787.
[42] S. Sawin, *Links, quantum groups and TQFTs*, Bull. Amer. Math. Soc. **33** (1996), 413–445.
[43] V.G. Turaev, *Quantum Invariants for Knots and 3-Manifolds*, de Gruyter, Berlin (1994).
[44] V.G. Turaev and O.Y. Viro, *State sum invariants of 3-manifolds and quantum 6j-symbols*, Topology **31** (1992), 865–902.
[45] V.G. Turaev and H. Wenzl, *Quantum invariants of 3-manifolds associated with classical simple Lie algebras*, Internat. J. Math. **4** (1993), 323–358.
[46] A.H. Wallace, *Modifications and cobounding manifolds*, Canad. J. Math. **12** (1960), 503–528.
[47] H. Wenzl, *On sequences of projections*, C. R. Math. Rep. Acad. Sci. **IX** (1987), 5–9.
[48] Y. Yokota, *On quantum* SU(2) *invariants and generalised bridge numbers of knots*, Math. Proc. Cambridge Philos. Soc. **117** (1995), 545–557.
[49] Y. Yokota, *Skeins and quantum* $\mathrm{SU}(N)$ *invariants of 3-manifolds*, Math. Ann. **307** (1997), 109–138.

CHAPTER 15

L^2-Invariants of Regular Coverings of Compact Manifolds and CW-Complexes

Wolfgang Lück

Fachbereich für Mathematik und Informatik, Westfälische Wilhelms-Universität Münster, Einsteinstr. 62, 48149 Münster, Germany

E-mail: lueck@math.uni-muenster.de

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Introduction

When we were asked to write this survey article for the Handbook of Geometry, we began to wonder who will read such an article. Well, certainly not someone who wants to fix his car. The possible readers could be (1) students or more advanced mathematicians who are looking for new problems, (2) people who are curious about this topic and want to get a first impression, (3) others who are interested in this topic already and want to invest some time to learn more about the techniques, (4) experts who want to see new developments and an extensive list of literature (with all their papers cited) or (5) people who want to read and absorb something new without big effort while watching a Star Trek episode or a soccer game. We hope that any of these groups can get something from this survey article. The appeal of the topic comes from its connections to rather different fields of mathematics like spectral theory of the Laplace operator, questions about metrics with certain curvature properties on manifolds, isomorphism conjectures in algebraic K-theory and the Baum–Connes Conjecture, low-dimensional manifolds, group theory, index theory, intersection homology, representation theory of Lie groups and so on.

For the groups (1) and (2) we recommend to glance at Section 1 and then read Section 2 where the main problems are stated. Moreover, there are other conjectures and open problems stated throughout the other sections. We treat Conjecture 2.1 about the rationality of the L^2-Betti numbers, the Singer Conjecture 2.6 about the vanishing of the L^2-Betti numbers outside the middle dimension of the universal covering of an aspherical closed manifold, the Hopf Conjecture 2.7 which is a version of the Singer Conjecture for closed Riemannian manifolds with negative sectional curvature, Conjecture 8.9 about the rationality and positivity of the Novikov–Shubin invariants, Conjecture 9.10 about the triviality of the homomorphism $\Phi_\Gamma : \mathrm{Wh}(\Gamma) \to \mathbb{R}^{>0}$ given by the Fuglede–Kadison determinant, Conjecture 9.18 about the relation of the simplicial volume and the L^2-torsion of aspherical closed manifolds, Conjecture 9.24 about the vanishing of the L^2-torsion for closed aspherical manifolds with amenable fundamental groups and the zero-in-the-spectrum Conjecture 11.1 and 11.4. None of these problems seem to be easy and have created a lot of work as discussed in Sections 3, 4, 7, 8 and 9. These sections and Sections 5, 6, 10 and 12 are completely independent of one another except for Section 9 which uses some information from Section 8, but nevertheless can be read without knowing Section 8. So the reader has not to be overwhelmed by the length of this article, but pick what he is interested in. For Section 11, however, we recommend to read through Section 8 first. For the groups (1), (2) and in particular (3) we have included some of the proofs in the text although they are very often somewhere in the literature. The motivation is sometimes that the proofs themselves are very illuminating or just nice, that the proofs are hard to find or that the proofs presented here are, hopefully, easier or better to understand than the one in the literature. It is certainly possible to read only the definitions, lemmas and theorems and skip all the proofs or additional information. We recommend this in particular to the fifth group of readers, and also the last episode of Star Trek – The Next Generation, we believe the title is "All good things" (in German TV "gestern, heute, morgen"). Although the list of references is quite long, it may well have happened that we have not cited a paper which should appear there, and we apologize for that.

This survey contains also mini-surveys on the following topics: 3-manifolds in Section 3, amenable groups in Section 4, residually finite groups in Section 5, Kähler manifolds in Section 7 and analytic Ray–Singer torsion and Reidemeister torsion in Section 9. There are other survey articles on topics of this survey article or related topics, for instance [42], [109, Section 8], [144,158,171,203,226,234].

We mention that we have for simplicity restricted ourselves to the von Neumann algebra of a group. One can formulate a lot of the results also for arbitrary von Neumann algebras. Group means always discrete group. We have restricted ourselves to regular coverings, applications of L^2-cohomology to other topics are very briefly mentioned in Section 12.
We wish to thank the Max-Planck-Institut in Bonn for its hospitality while parts of this article were written, John Lott for a lot of discussions about this topic and Clemens Bratzler for reading through the manuscript.

Münster,
Wolfgang Lück, February 1999

1. L^2-Betti numbers for CW-complexes of finite type

In this section we introduce L^2-Betti numbers for regular coverings of CW-complexes of finite type and discuss their main properties.

Let X be a connected CW-complex of finite type. *Finite type* means that each skeleton of X is finite, but X may have infinite dimension. The *pth Betti number* $b_p(X)$ is defined by the rank of the finitely generated abelian group $H_p(X)$ given by the cellular or, equivalently, singular homology. In algebraic topology it has turned out to be useful to improve classical invariants such as Euler characteristic and signature by passing to the universal covering and taking the action of the fundamental group into account. This leads, for instance, in the case of the Euler characteristic to the finiteness obstruction, and in the case of the signature to surgery obstructions such as symmetric signatures. We want to apply this strategy to Betti numbers.

The following naive approach does not work. One could think of applying an appropriate notion of dimension for modules over the integral group ring $\mathbb{Z}\pi$ of the fundamental group π to the homology $H_p(\widetilde{X})$ of the universal covering. Notice that the π-action on $\widetilde{X}$ induces a $\mathbb{Z}\pi$-module structure on $H_p(\widetilde{X})$. The problem is that in general $\mathbb{Z}\pi$ is not Noetherian. Hence $H_p(\widetilde{X})$ is not necessarily finitely generated although X is of finite type and therefore the pth module $C_p(\widetilde{X})$ of the cellular $\mathbb{Z}\pi$-chain complex is finitely generated over $\mathbb{Z}\pi$. So the dimension may be infinite. Moreover, it is not at all clear that the dimension is additive.

The reason why a lot of algebraic manipulations work nicely for the complex group ring of a finite group is that this ring is semi-simple. It has this property because there are enough projections, this being a consequence of the fact that $\mathbb{C}\Gamma$ is a Hilbert space. Namely, an ideal I in $\mathbb{C}\Gamma$ is a direct summand because it has an orthogonal complement. This property does not hold for infinite groups. However, if one enlarges the complex group ring to the von Neumann algebra, all the convenient properties of the complex group ring of a finite group carry over to arbitrary groups. This motivates the following definitions.

If Γ is a group, define $l^2(\Gamma)$ by the Hilbert space of square-summable formal sums $\sum_{\gamma\in\Gamma}\lambda_\gamma\gamma$ with complex coefficients λ_γ. Square-summable, of course, means

$\sum_{\gamma\in\Gamma}|\lambda_\gamma|^2<\infty$ and the inner product is given by

$$\Big\langle\sum_{\gamma\in\Gamma}\lambda_\gamma\gamma,\ \sum_{\gamma\in\Gamma}\mu_\gamma\gamma\Big\rangle=\sum_{\gamma\in\Gamma}\overline{\lambda_\gamma}\cdot\mu_\gamma.$$

The *group von Neumann algebra* of Γ is defined by the space of Γ-equivariant bounded operators from $l^2(\Gamma)$ to itself

$$\mathcal{N}(\Gamma)=\mathcal{B}\big(l^2(\Gamma),l^2(\Gamma)\big)^{\Gamma}, \tag{1.1}$$

where $l^2(\Gamma)$ is equipped with the obvious left Γ-action. This is not the standard definition, but equivalent to it. The *standard trace* of the von Neumann algebra is defined by

$$\mathrm{tr}=\mathrm{tr}_{\mathcal{N}(\Gamma)}:\mathcal{N}(\Gamma)\to\mathbb{C},\qquad f\mapsto\langle f(e),e\rangle_{l^{(2)}(\Gamma)},$$

where $e\in\Gamma\subset l^2(\Gamma)$ is the unit element. This trace extends to matrices

$$\mathrm{tr}:M\big(n,n,\mathcal{N}(\Gamma)\big)\to\mathbb{C} \tag{1.2}$$

by sending a matrix to the sum of the traces of the diagonal entries.

A *Hilbert $\mathcal{N}(\Gamma)$-module* is a Hilbert space V together with a left action of Γ by linear isometries such that there is a Hilbert space H and a Γ-equivariant isometric embedding of V into the tensor product of Hilbert spaces $l^2(\Gamma)\widehat{\otimes}H$. The isometric embedding is not part of the structure, only its existence is required. An example is $l^2(\Gamma)$ itself with the obvious left Γ-action. A Hilbert $\mathcal{N}(\Gamma)$-module is *finitely generated* if there is a surjective bounded Γ-equivariant operator from $l^2(\Gamma)^n=\bigoplus_{i=1}^n l^2(\Gamma)$ onto V for an appropriate positive integer n. This is equivalent to the existence of an isometric Γ-equivariant embedding of V into $l^2(\Gamma)^n$ for an appropriate positive integer n and to the existence of an orthogonal Γ-equivariant projection $\mathrm{pr}:l^2(\Gamma)^n\to l^2(\Gamma)^n$ whose image is isometrically Γ-isomorphic to V for an appropriate positive integer n.

A *morphism of Hilbert $\mathcal{N}(\Gamma)$-modules* $f:U\to V$ is a bounded Γ-equivariant operator. We call f a *weak isomorphism* if its kernel is trivial and its image is dense. The polar decomposition of a weak isomorphism f looks like $i\circ|f|$ where $|f|:U\to U$ is a positive morphism and $i:U\to V$ an isometric isomorphism. In particular U and V are isometrically isomorphic if there is a weak isomorphism from U to V. A sequence of Hilbert $\mathcal{N}(\Gamma)$-modules is called *exact* if i is injective, $\mathrm{im}(i)=\ker(p)$ and p is surjective. It is called *weakly exact* if i is injective, $\mathrm{clos}(\mathrm{im}(i))=\ker(p)$ and $\mathrm{clos}(\mathrm{im}(p))=W$.

DEFINITION 1.3. Let V be a finitely generated Hilbert $\mathcal{N}(\Gamma)$-module. Define its *von Neumann dimension* by

$$\dim(V)=\dim_{\mathcal{N}(\Gamma)}(V)=\mathrm{tr}(\mathrm{pr})\in[0,\infty),$$

where $\mathrm{pr}:l^2(\Gamma)^n\to l^2(\Gamma)^n$ is any orthogonal Γ-equivariant projection whose image is isometrically Γ-isomorphic to V.

It is not hard to check that the definition above is independent of the choice of projection. The elementary proof of the next result is left to the reader. It follows from the general properties of the universal center valued trace of a finite von Neumann algebra [131, Theorem 8.2.8, p. 517, Proposition 8.3.10, p. 525 and Theorem 8.4.3, p. 532].

LEMMA 1.4. *Let U, V and W be finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Then*

(1) *Faithfulness*: $\dim_{\mathcal{N}(\Gamma)}(U) = 0$ *if and only if* $U = 0$;

(2) *Monotony: If* $U \subset V$ *then*

$$\dim_{\mathcal{N}(\Gamma)}(U) \leqslant \dim_{\mathcal{N}(\Gamma)}(V);$$

(3) *Continuity: If* $U_1 \supset U_2 \supset \cdots$ *is a nested sequence of Hilbert* $\mathcal{N}(\Gamma)$*-submodules of* U*, then*

$$\dim_{\mathcal{N}(\Gamma)}\left(\bigcap_{n=1}^{\infty} U_n\right) = \lim_{n\to\infty} \dim_{\mathcal{N}(\Gamma)}(U_n);$$

(4) *Weak exactness: If* $0 \to U \xrightarrow{j} V \xrightarrow{q} W \to 0$ *is weakly exact, then*

$$\dim_{\mathcal{N}(\Gamma)}(V) = \dim_{\mathcal{N}(\Gamma)}(U) + \dim_{\mathcal{N}(\Gamma)}(W).$$

We will extend this notion to arbitrary modules over $\mathcal{N}(\Gamma)$ in Section 10.

DEFINITION 1.5. Let X be a (not necessarily connected) CW-complex of finite type. Let $p: \overline{X} \to X$ be a regular covering of X with group of deck transformations Γ acting from the left. Define the *cellular L^2-chain complex* of $\overline{X}$ by the chain complex of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules

$$C_*^{(2)}(\overline{X}) = l^2(\Gamma) \otimes_{\mathbb{Z}\Gamma} C_*(\overline{X}).$$

The *cellular L^2-cochain complex* is defined by

$$C_{(2)}^*(\overline{X}) = \hom_{\mathbb{Z}\Gamma}(C_*(\overline{X}), l^2(\Gamma))$$

where for $\gamma \in \Gamma$ and $f \in C_{(2)}^*(\overline{X})$ the element $\gamma \cdot f$ sends $x \in C_*(\overline{X})$ to $f(x)\gamma^{-1}$. Define the *L^2-homology* of $\overline{X}$ by

$$H_p^{(2)}(\overline{X}) = \ker(c_p^{(2)}) / \operatorname{clos}(\operatorname{im}(c_{p+1}^{(2)})),$$

where $c_*^{(2)}$ is the differential of $C_*^{(2)}(\overline{X})$. The definition of *L^2-cohomology* is analogous. Define the *pth L^2-Betti number* of $\overline{X}$ by

$$b_p^{(2)}(\overline{X}) = \dim_{\mathcal{N}(\Gamma)}(H_p^{(2)}(\overline{X})) = \dim_{\mathcal{N}(\Gamma)}(H^p_{(2)}(\overline{X})).$$

The decisive difference between L^2-homology and the Γ-equivariant homology with coefficients in $l^2(\Gamma)$ viewed as $\mathbb{Z}\Gamma$-module is that we divide by the closure of the image of the corresponding differential and not only by the image itself. The reason is that in the L^2-setting we want to keep the Hilbert space structure coming from the cellular L^2-chain complex. In order to guarantee completeness we must divide by a closed subspace. We will investigate the difference between these two homologies later when we introduce Novikov–Shubin invariants in Section 8 which measure this difference. Hilbert complexes in general (not necessarily over a finite von Neumann algebra) are treated in [39].

If we deal with a smooth manifold M, then these definitions are understood for the CW-complex structure given by some smooth triangulation. We will prove in Theorem 1.7 that the choice of triangulation does not matter because two triangulations of M give homotopy equivalent CW-complexes. In the sequel we will denote by $\widetilde{X}$ the universal covering with the fundamental group $\pi_1(X)$ as group of deck transformations Γ, provided that X is connected.

Notice that both $H_p^{(2)}(\overline{X})$ and $H^p_{(2)}(\overline{X})$ are isometrically Γ-isomorphic to the kernel of the *combinatorial Laplace operator*

$$\Delta_p = \left(c_p^{(2)}\right)^* \circ c_p^{(2)} + c_{p+1}^{(2)} \circ \left(c_{p+1}^{(2)}\right)^* : C_p^{(2)}\left(\overline{X}\right) \to C_p^{(2)}\left(\overline{X}\right). \tag{1.6}$$

The elementary proof can be found in [163, Theorem 3.7, p. 230]. Hence there is no difference between homology and cohomology in the L^2-setting.

Next we discuss the main properties of L^2-Betti numbers, in particular in comparison with the ones of the ordinary Betti numbers.

THEOREM 1.7.

(1) *Homotopy invariance*
Let $\overline{X}$ and $\overline{Y}$ be regular coverings of CW-complexes X and Y of finite type with the same group Γ of deck transformations. Let $f : \overline{X} \to \overline{Y}$ be a Γ-equivariant map. If f is a homotopy equivalence, then

$$b_p^{(2)}\left(\overline{X}\right) = b_p^{(2)}\left(\overline{Y}\right) \quad \textit{for } 0 \leqslant p.$$

If f is d-connected, i.e., f induces an isomorphism on π_n for $n < d$ and an epimorphism on π_d, then

$$b_p^{(2)}\left(\overline{X}\right) = b_p^{(2)}\left(\overline{Y}\right) \quad \textit{for } p < d;$$
$$b_d^{(2)}\left(\overline{X}\right) \geqslant b_d^{(2)}\left(\overline{Y}\right).$$

(2) *Euler–Poincaré formula*
Let $\overline{X}$ be a regular covering of a finite CW-complex X. Let

$$\chi(X) = \sum_{p \geqslant 0} (-1)^p \cdot \beta_p(X) \in \mathbb{Z}$$

be the Euler characteristic of X where $\beta_p(X)$ is the number of p-cells of X. Then

$$\chi(X) = \sum_{p\geqslant 0} (-1)^p \cdot b_p^{(2)}(\overline{X});$$

(3) *Poincaré duality*
Let $\overline{M}$ be a regular covering of the closed manifold M of dimension n. Then

$$b_p^{(2)}(\overline{M}) = b_{n-p}^{(2)}(\overline{M});$$

(4) *Künneth formula*
Let X and Y be CW-complexes of finite type. Let $\overline{X}$ and $\overline{Y}$ be regular coverings of X and Y. Then $\overline{X} \times \overline{Y}$ is a regular covering of $X \times Y$ and

$$b_n^{(2)}(\overline{X} \times \overline{Y}) = \sum_{p+q=n} b_p^{(2)}(\overline{X}) \cdot b_q^{(2)}(\overline{Y}) \quad \textit{for } n \geqslant 0;$$

(5) *Morse inequalities*
Let $\overline{X}$ be a regular covering of a CW-complex X of finite type. Let $\beta(X)$ be the number of p-cells in X. Then

$$\sum_{p=0}^{n} (-1)^{n-p} \cdot b_p^{(2)}(\overline{X}) \leqslant \sum_{p=0}^{n} (-1)^{n-p} \cdot \beta_p(X) \quad \textit{for } n \geqslant 0;$$

(6) *L^2-Hodge–deRham decomposition*
Let $\overline{M}$ be a covering of the oriented closed Riemannian manifold M with deck transformation group Γ. Let $\mathcal{H}_{(2)}^p(\overline{M})$ be the space of harmonic smooth L^2-p-forms on $\overline{M}$, i.e., smooth p-forms ω on $\overline{M}$ such that $\int_{\overline{M}} \omega \wedge *\omega$ is finite and ω lies in the kernel of the Laplace operator with respect to the induced Riemannian metric on $\overline{M}$. Then integration defines an isomorphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules

$$\mathcal{H}_{(2)}^p(\overline{M}) \to H_{(2)}^p(\overline{M});$$

(7) *Multiplicative property for finite coverings*
Let X be a CW-complex of finite type and $p : \overline{X} \to X$ be a regular covering with group of deck transformations Γ. Let $\Gamma_0 \subset \Gamma$ be a subgroup of Γ of finite index n. We obtain a regular covering denoted by $\overline{\overline{X}}$ by $\overline{X} \to \overline{X}/\Gamma_0$. Notice that the coverings $\overline{\overline{X}}$ and $\overline{X}$ have the same total spaces but different groups of deck transformations. Then

$$b_p^{(2)}(\overline{\overline{X}}) = n \cdot b_p^{(2)}(\overline{X}) \quad \textit{for } p \geqslant 0;$$

(8) *L^2-Betti numbers for finite groups Γ*
Let X be a CW-complex of finite type and let $p:\overline{X}\to X$ be a regular covering with group of deck transformations Γ of finite order $|\Gamma|$. Then:

$$b_p^{(2)}(\overline{X}) = \frac{1}{|\Gamma|}\cdot b_p(\overline{X}) \quad \textit{for } p \geqslant 0;$$

(9) *Zero-th L^2-Betti number*
Let X be a connected CW-complex of finite type and let $p:\overline{X}\to X$ be a regular covering with group of deck transformations Γ and connected $\overline{X}$. Then

$$b_0^{(2)}(\overline{X}) = \begin{cases} \frac{1}{|\Gamma|} & \textit{if } |\Gamma| < \infty; \\ 0 & \textit{otherwise}; \end{cases}$$

(10) *S^1-actions and L^2-Betti numbers*
Let M be a connected closed manifold with S^1-action. Suppose that for one orbit S^1/H (and hence for all orbits) the inclusion into M induces a map on π_1 with infinite image. (In particular the S^1-action has no fixed points.) Then

$$b_p^{(2)}(\widetilde{M}) = 0 \quad \textit{for } p \geqslant 0.$$

PROOF. (1) Because of the Equivariant Cellular Approximation Theorem we can assume that f is cellular. Since $\overline{X}$ and $\overline{Y}$ are free Γ-CW-complexes the map f is even a Γ-homotopy equivalence. Because of the Equivariant Cellular Approximation Theorem there is a cellular Γ-map $g:\overline{Y}\to\overline{X}$ such that there is a cellular Γ-homotopy between $f\circ g$ (respectively $g\circ f$) and the identity. One easily checks that two cellular Γ-maps which are connected by a cellular Γ-homotopy induce the same map on L^2-homology. Now the claim for a homotopy equivalence f follows. The more general case of a d-connected Γ-map f is proven in [148, Lemma 3.3].

(2) This follows as in the classical situation, where $\chi(X)$ is expressed in terms of (ordinary) Betti numbers, from the fact that the von Neumann dimension is weakly additive (see Theorem 1.4.4).

(3) Follows from the Poincaré $\mathbb{Z}\Gamma$-chain homotopy equivalence [247, Theorem 2.1, p. 23]

$$\bigcap[M]: C^{n-*}(\overline{M}) \to C_*(\overline{M}).$$

(4) Follows from the isomorphism of cellular chain complexes

$$C(\overline{X})\otimes_{\mathbb{Z}} C(\overline{Y}) \to C(\overline{X}\times\overline{Y}).$$

See also [255, Corollary 2.36, p. 181].

(5) Analogous to (2).

(6) Is proven by Dodziuk [72].

(7) If V is a finitely generated $\mathcal{N}(\Gamma)$-module and $\operatorname{res}(V)$ its restriction to $\mathcal{N}(\Gamma_0)$ which is a finitely generated Hilbert $\mathcal{N}(\Gamma_0)$-module, then:

$$\dim_{\mathcal{N}(\Gamma_0)}\bigl(\operatorname{res}(V)\bigr) = n \cdot \dim_{\mathcal{N}(\Gamma)}(V).$$

(8) Follows from (7).

(9) If Γ has finite order, this follows from (8). If Γ is infinite, this follows from the fact that $l^2(\Gamma)^\Gamma$ is trivial.

(10) is proven in [163, Theorem 3.20, p. 235]. □

More information about Morse inequalities and L^2-invariants can be found in [89,90, 157,172,173,196,197,234].

EXAMPLE 1.8. A good source of well understood examples is given by the special case where Γ is the free abelian group $\mathbb{Z}^r$ of rank r. On one hand everything becomes simple, on the other hand one can already see some of the important phenomenons in this special case. See also [57, Section 5], [76].

One simplification comes from Fourier transformation. Namely, we obtain a natural isometric $\mathbb{Z}^r$-equivariant isomorphism

$$l^2(\mathbb{Z}^r) \to L^2(T^r),$$

where $L^2(T^r)$ is the Hilbert space of L^2-functions on T^r. Let $L^\infty(T^r)$ be the C^*-algebra of essentially bounded measurable functions on T^r. Then we obtain an isomorphism of C^*-algebras

$$M : L^\infty(T^r) \to \mathcal{N}(\mathbb{Z}^r) = \mathcal{B}\bigl(L^2(T^r), L^2(T^r)\bigr)^{\mathbb{Z}^r} \quad f \mapsto M_f,$$

where $M_f : L^2(T^r) \to L^2(T^r)$ sends g to the function $f \cdot g$ which assigns to $z \in T^r$ the element $f(z) \cdot g(z)$. A morphism of Hilbert $\mathcal{N}(\mathbb{Z}^r)$-modules $L^2(T^r) \to L^2(T^r)$ is given by M_f for some $f \in L^\infty(T^r)$. It is a weak isomorphism if and only if $f^{-1}(0)$ is a set of measure zero, and it is an isomorphism if for some $\varepsilon > 0$ the set $\{z \in T^r \mid |f(z)| < \varepsilon\}$ has measure zero. In particular we see concrete examples of weak isomorphisms which are not isomorphisms. An important example is given by $M_{z_i - 1}$. The operator M_f is positive if and only if f takes values in the non-negative real numbers. In this case the spectral family $\{E_\lambda \mid \lambda \in \mathbb{R}\}$ is given by $E_\lambda = M_{\chi_\lambda}$ where χ_λ is the characteristic function of the set $\{z \in T^r \mid f(z) \leqslant \lambda\}$. The von Neumann trace $\operatorname{tr}_{\mathcal{N}(\mathbb{Z}^r)}$ becomes

$$\operatorname{tr}_{\mathcal{N}(\mathbb{Z}^r)} : L^\infty(T^r) \to \mathbb{C}, \quad f \mapsto \int_{T^r} f.$$

Let $\overline{X} \to X$ be a regular covering of a CW-complex of finite type with $\mathbb{Z}^r$ as group of deck transformations. Let $\mathbb{Z}[\mathbb{Z}^r]_{(0)}$ be the quotient field of the integral group ring of $\mathbb{Z}^r$.

Let $\dim_{\mathbb{Z}[\mathbb{Z}^r]_{(0)}}(H_p(\overline{X}) \otimes_{\mathbb{Z}[\mathbb{Z}^r]} \mathbb{Z}[\mathbb{Z}^r]_{(0)})$ be the dimension of the finite-dimensional vector space of the quotient field. Then we get [157, Example 4.3]

$$b_p^{(2)}(\overline{X}) = \dim_{\mathbb{Z}[\mathbb{Z}^r]_{(0)}}\big(H_p(\overline{X}) \otimes_{\mathbb{Z}[\mathbb{Z}^r]} \mathbb{Z}[\mathbb{Z}^r]_{(0)}\big).$$

REMARK 1.9. The L^2-Hodge–deRham decomposition of Theorem 1.4.6 proves that for a regular covering $\overline{M} \to M$ of a closed Riemannian manifold M the L^2-Betti numbers have the following analytic interpretation. Let $L^2\Omega^p(\overline{M})$ be the Hilbert space of all square-integrable $\mathbb{C}$-valued p-forms on $\overline{M}$. This is the Hilbert space completion of the space $C_0^\infty\Omega^p(\overline{M})$ of smooth $\mathbb{C}$-valued p-forms on $\overline{M}$ with compact support and the standard L^2-pre-Hilbert structure. Since $\overline{M}$ is complete (with respect to the lifted Riemannian metric), the Laplace operator Δ_p is essentially selfadjoint in $L^2\Omega^p(\overline{M})$, i.e., its closure with respect to the domain $C_0^\infty\Omega^p(\overline{M})$ is a self-adjoint operator on $L^2\Omega^p(\overline{M})$ [55]. Let $\Delta_p = \int \lambda \, dE_\lambda^p$ be the spectral decomposition with right-continuous spectral family $\{E_\lambda^p \mid \lambda \in \mathbb{R}\}$. Let $E_\lambda^p(\overline{x}, \overline{y})$ be the Schwartz kernel of E_λ^p. Since $E_\lambda^p(\overline{x}, \overline{x})$ is an endomorphism of a finite-dimensional complex vector space, its trace $\mathrm{tr}_{\mathbb{C}}$ is defined. Let $\mathcal{F}$ be a fundamental domain for the Γ-action on $\overline{M}$, i.e., an open subset $\mathcal{F}$ in $\overline{M}$ such that $\bigcup_{\gamma\in\Gamma} \gamma \operatorname{clos}(\mathcal{F}) = \overline{M}$ and $\gamma\mathcal{F} \cap \mathcal{F} = \emptyset$ for $\gamma \in \Gamma$ with $\gamma \neq 1$ [210, Section 6.5]. Then the *analytic pth spectral density function* is defined by

$$F^p(\lambda) = \int_{\mathcal{F}} \mathrm{tr}_{\mathbb{C}}\big(E_\lambda^p(\overline{x}, \overline{x})\big) \, d\overline{x}, \quad \lambda \in \mathbb{R}. \tag{1.10}$$

We will later investigate this spectral density function more closely. We mention the equality

$$b_p^{(2)}(\overline{M}) = F^p(0). \tag{1.11}$$

By means of a Laplace transformation we obtain the equality

$$b_p^{(2)}(\overline{M}) = \lim_{t\to\infty} \int_{\mathcal{F}} \mathrm{tr}_{\mathbb{C}}\big(e^{-t\Delta_p}(\overline{x}, \overline{x})\big) \, d\overline{x}, \tag{1.12}$$

where $e^{-t\Delta_p}(\overline{x}, \overline{y})$ denotes the heat kernel on $\overline{M}$, i.e., the Schwartz kernel of $e^{-t\Delta_p}$.

All the definitions and results above extend to pairs of proper cocompact Γ-CW-complexes and manifolds with boundary and proper cocompact Γ-actions [163, Sections 3 and 5]. The L^2-Hodge–deRham decomposition for manifolds with boundary is proven in [228].

2. Basic conjectures

In this section we state the (in our view) most important conjectures concerning L^2-Betti numbers. They motivate a lot of the work done on this subject. We also give a list of theorems which give evidence for them. Results of Sections 3, 4 and 7 also prove these conjectures in special cases. More questions and conjectures will be discussed in other sections.

CONJECTURE 2.1 (*Rationality of L^2-Betti numbers*). Let Γ be a group and A be a (m, n)-matrix with coefficients in $\mathbb{Z}\Gamma$. Let $f: l^2(\Gamma)^m \to l^2(\Gamma)^n$ be the Γ-equivariant bounded operator induced by right multiplication with A. Then

(1) $\dim_{\mathcal{N}(\Gamma)}(\ker(f))$ is rational;
(2) Let k be a positive integer such that the order of any finite subgroup of Γ divides k. Then $k \cdot \dim_{\mathcal{N}(\Gamma)}(\ker(f))$ is an integer;
(3) If Γ is torsionfree, then $\dim_{\mathcal{N}(\Gamma)}(\ker(f))$ is an integer.

Conjecture 2.1 above has the following equivalent reformulations provided that Γ is finitely presented.

LEMMA 2.2. *The following statements are equivalent for a finitely presented group* Γ:

(1) Γ *satisfies Conjecture* 2.1.1 (*respectively* 2.1.2, *respectively* 2.1.3);
(2) *If* X *is a connected CW-complex of finite type with* Γ *as fundamental group, then* $b_p^{(2)}(\widetilde{X})$ *satisfies the corresponding statement of Conjecture* 2.1.1 (*respectively* 2.1.2, *respectively* 2.1.3) *for all* $p \geqslant 0$;
(3) *If* M *is a closed manifold with* Γ *as fundamental group, then* $b_p^{(2)}(\widetilde{M})$ *satisfies the corresponding statement of Conjecture* 2.1.1 (*respectively* 2.1.2, *respectively* 2.1.3) *for all* $p \geqslant 0$.

PROOF. Obviously (1) implies (2) and (2) implies (3) so that it remains to prove that (3) implies (1). The key observation is the following. Let A be a (m, n)-matrix over $\mathbb{Z}\Gamma$ and $d \geqslant 3$ be an integer. Choose a finite connected 2-dimensional CW-complex X with Γ as fundamental group. By attaching n $(d-1)$-cells to X with a trivial attaching map and m d-cells one can construct a finite CW-complex Y such that the dth differential of the cellular $\mathbb{Z}\Gamma$-chain complex $C(\widetilde{Y})$ is, for $d \geqslant 4$, given by A and for $d = 3$ by the map induced by A composed with the canonical inclusion $l^2(\Gamma)^n \to l^2(\Gamma)^n \oplus l^2(\Gamma)^{n'}$ where n' is the number of 2-cells in X. Let $f: l^2(\Gamma)^m \to l^2(\Gamma)^n$ be the bounded Γ-equivariant operator induced by A. Since Y has no $(d+1)$-cells, the kernel of f is just $H_d^{(2)}(\widetilde{Y})$. Next one can embed the finite d-dimensional CW-complex Y into $\mathbb{R}^{2d+2}$. Let M be the boundary of a regular neighborhood of X [221, Chapter 3]. Then there is a $(d+1)$-connected map $M \to X$ and because of Theorem 1.7.1

$$b_d^{(2)}(\widetilde{M}) = b_d^{(2)}(\widetilde{Y}) = \dim(\ker(f)). \qquad \square$$

REMARK 2.3. If Conjecture 2.1 holds for all finitely generated groups it holds for all groups. Namely, given a (m, n) matrix A with coefficients in $\mathbb{Z}\Gamma$, it suffices to look at the subgroup Γ' generated by those elements which appear in one of the entries of A with non-zero coefficients. Namely, A can be viewed as a matrix over $\mathbb{Z}\Gamma'$ and the dimension over Γ of the associated Γ-equivariant bounded operator agrees with the one over Γ'.

Lemma 2.2 remains true if we substitute the assumption finitely presented for Γ by finitely generated and substitute the universal coverings $\widetilde{X}$ (respectively $\widetilde{M}$) by regular coverings $\overline{X} \to X$ (respectively $\overline{M} \to M$) for a connected finite CW-complex (respectively closed manifold M). The modification in the proof is as follows. Choose a finitely generated free group F together with an epimorphism $p: F \to \Gamma$. Lift A to a matrix A' over $\mathbb{Z}[F]$ and construct X, Y and M as explained in the proof of Lemma 2.2 for F and A'. Then take the coverings of Y (respectively M) associated to p with Γ as group of deck transformations.

Hence Conjecture 2.1 is true for all groups Γ and matrices A if and only if for all regular coverings $\overline{M} \to M$ of closed manifolds M with a finitely generated group of deck transformations $b_p^{(2)}(\overline{M})$ satisfies the corresponding statement.

The question of whether the third statement in Lemma 2.2 is true was asked by Atiyah [3, p. 72]. Next we show that Conjecture 2.1 implies the Zero-Divisor-Conjecture. For a discussion of the Zero-Divisor-Conjecture and related conjectures we refer to [138, p. 95].

LEMMA 2.4. *If the group Γ satisfies Conjecture* 2.1.3, *then it also satisfies the Zero-Divisor-Conjecture which says: The rational group ring $\mathbb{Q}\Gamma$ has no non-trivial zero-divisors if and only if Γ is torsionfree.*

PROOF. Suppose that Γ is not torsionfree. Let $g \in \Gamma$ be an element of finite order $|g| \neq 1$. Then the norm element $N = \sum_{k=1}^{|g|} g^k$ is a zero-divisor in $\mathbb{Q}\Gamma$ since it satisfies

$$N \cdot (|g| - N) = 0.$$

Suppose that Γ is torsionfree. Let $x \in \mathbb{Q}\Gamma$ be a zero-divisor. We have to show, under the assumption that Conjecture 2.1.3 is true, that x is trivial. By multiplying x with an appropriate integer, we can assume without loss of generality that x belongs to $\mathbb{Z}\Gamma$. Right multiplication with x induces a Γ-equivariant bounded operator $r_x : l^2(\Gamma) \to l^2(\Gamma)$. Conjecture 2.1 implies that $\dim_{\mathcal{N}(\Gamma)}(\ker(r_x))$ is an integer. As x is a zero-divisor, the kernel of r_x is not trivial and hence $\dim_{\mathcal{N}(\Gamma)}(\ker(r_x))$ is not zero. Since $\dim_{\mathcal{N}(\Gamma)}(\ker(r_x))$ is bounded by the dimension of $l^2(\Gamma)$, which is 1, we get $\dim_{\mathcal{N}(\Gamma)}(\ker(r_x)) = 1 = \dim_{\mathcal{N}(\Gamma)}(l^2(\Gamma))$. As the kernel of r_x is a closed subspace of $l^2(\Gamma)$, this implies that r_x is the trivial map. Hence x is zero. This shows that the rational group ring has no non-trivial zero-divisors. □

Let $\mathcal{C}$ be the smallest class of groups which (i) contains all free groups, (ii) is closed under directed unions and (iii) satisfies $G \in \mathcal{C}$ whenever G contains a normal subgroup H such that H belongs to $\mathcal{C}$ and G/H is elementary amenable. We recall that the class of *elementary amenable* groups is defined as the smallest class of groups which contains all finite and all abelian groups, and is closed under taking subgroups, forming factor groups, group extensions, and upwards directed unions.

THEOREM 2.5 (Linnell [140]). *Conjecture* 2.1 *holds for the class $\mathcal{C}$ of groups.*

The key result in [140] is that for a torsionfree group Γ in the class $\mathcal{C}$ there is a division ring $D(\Gamma)$ satisfying $\mathbb{C}\Gamma \subset D(\Gamma) \subset U(\Gamma)$, where $U(\Gamma)$ is the algebra of closed densely defined operators $l^2(\Gamma) \to l^2(\Gamma)$ which are affiliated to the group von Neumann algebra. Linnell uses the Fredholm technique developed by Connes for his proof that the reduced C^*-algebra of a free group has no non-trivial projections [62, Section 7], [83], Moody's induction theorem [178, Theorem 1] and Cohn's theory of universal fields of fractions [61]. This indicates that there must be a connection between Conjecture 2.1 and the Baum–Connes Conjecture for the topological K-theory of the reduced C^*-algebra of a group [13, Conjecture 3.15, p. 254] and the Isomorphisms Conjecture for the algebraic K-theory of the integral group ring of Farrell and Jones [93, 1.6, p. 257]. See [68] for a unified treatment of the last two conjectures and see [213] for more details on this connection and a general strategy based on Linnell's work how to approach Conjecture 2.1. The Baum–Connes Conjecture says that one can compute the topological K-theory of the reduced C^*-algebra of a group Γ by a complicated induction process from the topological K-theory of the complex group rings of all finite subgroups of Γ. Conjecture 2.1 can be interpreted similarly, namely all the possible dimensions of kernels of Γ-equivariant bounded operators $l^2(\Gamma)^m \to l^2(\Gamma)^n$ which are induced by matrices over $\mathbb{Z}\Gamma$ are coming from the dimensions of kernels of Γ-equivariant bounded operators $l^2(\Gamma)^m \to l^2(\Gamma)^n$ which are induced by matrices over $\mathbb{Z}G$ for all finite subgroups $G \subset \Gamma$. Notice that the dimension of such operators coming from a finite group G are rational numbers which become integral when multiplied with the order of G. The missing link seems to be the not at all understood passage from finitely presented $\mathbb{Z}\Gamma$-modules to modules over the C^*-algebra of Γ. The connection between the generalization of the Euler Poincaré formula for finitely dominated CW-complexes and the Bass Conjecture [12, p. 156] which is related to the Baum–Connes Conjecture and the Isomorphisms Conjecture for the algebraic K-theory is treated by Eckmann [82].

CONJECTURE 2.6 (*Singer Conjecture*). If M is a closed aspherical Riemannian manifold of dimension n, then

$$b_p^{(2)}(\widetilde{M}) = 0 \quad \text{for } 2p \neq n.$$

If additionally $n = 2d$, then

$$(-1)^d \cdot \chi(M) \geqslant 0.$$

The conjecture above was originally only formulated for closed Riemannian manifolds with non-positive sectional curvature. Notice that any such manifold is aspherical by Hadamard's Theorem [99, 3.87, p. 134]. We recall that a space X is *aspherical* if its universal covering is contractible, or equivalently, all the higher homotopy groups of X are trivial. We will see that the Singer Conjecture 2.6 is true if $\pi_1(M)$ contains a normal non-trivial amenable subgroup in Section 4, and if M has dimension 3 and is not exceptional in Section 3. The Singer Conjecture 2.6 is true if $\widetilde{M}$ is a symmetric space of non-compact type. It is also true if M carries a non-trivial S^1-action, because then the inclusion of an orbit into M induces a map on the fundamental groups with infinite image [66, Lemma 5.1, p. 242 and Corollary 5.3, p. 243] and Theorem 1.7.10 applies.

CONJECTURE 2.7 (*Hopf Conjecture*). If M is a closed $2d$-dimensional Riemannian manifold of negative sectional curvature, then

$$b_d^{(2)}(\widetilde{M}) > 0; \qquad (-1)^d \cdot \chi(M) > 0.$$

In the Hopf Conjecture 2.7 above, the part about the Euler characteristic goes back to Hopf. The statements about the Euler characteristic in Conjecture 2.6 and 2.7 follow from the one about the L^2-Betti numbers by the Euler–Poincaré formula $\chi(M) = \sum_{p\geqslant 0}(-1)^p \cdot b_p^{(2)}(\widetilde{M})$ of Theorem 1.7.2.

Conjecture 2.7 has been proven by Dodziuk for a hyperbolic closed Riemannian manifold. Namely, we have already mentioned in Theorem 1.7 that there is L^2-Hodge–deRham decomposition [72]. Since the universal covering of a closed n-dimensional hyperbolic manifold is isometrically isomorphic to the n-dimensional hyperbolic space $\mathbb{H}^n$, and the von Neumann dimension of a finitely generated Hilbert $\mathcal{N}(\Gamma)$-module is zero if and only if the module is zero, the pth L^2-Betti number of M is trivial if and only if the space of harmonic L^2-integrable p-forms on $\mathbb{H}^n$ is trivial. This space is computed in [73] using the rotational symmetry of $\mathbb{H}^n$. Donnelly and Xavier [79] have proven for a closed n-dimensional Riemannian manifold M whose sectional curvature is pinched between -1 and D_n for some real number $-1 \leqslant D_n < -\frac{(n-2)^2}{(n-1)^2}$ that

$$b_p^{(2)}(\widetilde{M}) = 0 \quad \text{for } p \neq \frac{n}{2}, \frac{n \pm 1}{2}.$$

In particular, the Singer Conjecture 2.6 is true for such a manifold if n is even. This result has been improved by Jost and Yuanlong [129]. They assume that the sectional curvature of M satisfies $-a^2 \leqslant K \leqslant 0$ and the Ricci curvature is bounded from above by $-b^2$ for positive constants a, b and show that

$$b_p^{(2)}(\widetilde{M}) = 0 \quad \text{for } p \neq \frac{n}{2}, \ 2pa \leqslant b.$$

We will see in Section 7 that the Hopf Conjecture 2.7 is true if M is a Kähler manifold. Without some finiteness conditions on $\widetilde{X}$, such as being the total space of a regular covering over a finite CW-complex, Conjecture 2.6 becomes false. This follows from the work of Anderson [2] where the non-vanishing of the (reduced) L^2-cohomology of perturbations of the hyperbolic metric on the hyperbolic space is proved.

3. Low-dimensional manifolds

In this section we give information about the L^2-Betti numbers of universal coverings of compact manifolds of dimension $\leqslant 3$. We will only consider orientable manifolds since one gets the general case by passing to the orientation covering and the multiplicative property of the L^2-Betti numbers of Theorem 1.7.7.

EXAMPLE 3.1. We begin with a compact connected 1-dimensional manifold M. If M has boundary, it is $[0,1]$ and hence homotopy equivalent to a point, and its L^2-Betti numbers agree with the (ordinary) Betti numbers of the one-point-space, i.e., they are trivial except the 0th one which is 1. If M has no boundary, then M is S^1. Since there is a covering $S^1 \to S^1$ of degree d for $d \geqslant 2$, we conclude from the multiplicative property in Theorem 1.7.7

$$b_p^{(2)}(\widetilde{S^1}) = 0 \quad \text{for } p \geqslant 0.$$

It is illuminating to compute $b_p^{(2)}(\widetilde{S^1})$ directly. One easily checks that the cellular L^2-chain complex is concentrated in dimension 0 and 1 and its first differential $l^2(\mathbb{Z}) \xrightarrow{t-1} l^2(\mathbb{Z})$ is given by multiplication with the element $t-1$ for t a generator of the fundamental group $\pi_1(S^1) = \mathbb{Z}$. Its Fourier transformation is the bounded $\mathbb{Z}$-equivariant operator

$$M_{z-1}: L^2(S^1) \to L^2(S^1), \qquad f(z) \mapsto (z-1) \cdot f(z),$$

where we regard S^1 as a subset of the complex numbers $\mathbb{C}$. Obviously the kernel of M_{z-1} and hence $H_1^{(2)}(\widetilde{S^1})$ is trivial. One easily checks that $H_0^{(2)}(\widetilde{S^1})$ is the kernel of M_{z-1}^*, which is the kernel of $M_{z^{-1}-1}$ and hence trivial.

EXAMPLE 3.2. Let F_g^d be the orientable closed surface of genus g with d embedded open 2-disks removed. From Theorem 1.7 and the fact that a compact surface with boundary is homotopy equivalent to a bouquet of circles, one derives

$$b_0^{(2)}(\widetilde{F_g^d}) = \begin{cases} 1 & g=0, d=0,1, \\ 0 & \text{otherwise}, \end{cases}$$

$$b_1^{(2)}(\widetilde{F_g^d}) = \begin{cases} 0 & g=0, d=0,1, \\ d+2(g-1) & \text{otherwise}, \end{cases}$$

$$b_2^{(2)}(\widetilde{F_g^d}) = \begin{cases} 1 & g=0, d=0, \\ 0 & \text{otherwise}. \end{cases}$$

In the sequel 3-manifold means a compact connected orientable 3-dimensional manifold. Notice that in dimension 3 there is no difference between topological, PL- or smooth manifolds [176,191]. We want to compute the L^2-Betti numbers of the universal covering of a 3-manifold. For this purpose we have to recall some basic facts about such manifolds which are interesting in their own right. For more information about 3-manifolds we refer to [116,231,240].

A 3-manifold M is *prime* if, given a decomposition $M_1 \sharp M_2$ of M as a connected sum, M_1 or M_2 is homeomorphic to S^3. It is *irreducible* if every embedded bicollared 2-sphere bounds an embedded 3-disk. Any prime 3-manifold is irreducible or is homeomorphic to $S^1 \times S^2$ [116, Lemma 3.13]. One can write a 3-manifold M as a connected sum

$$M = M_1 \sharp M_2 \sharp \cdots \sharp M_r,$$

where each M_j is prime, and this prime decomposition is unique up to renumbering and oriented homeomorphism [116, Theorems 3.15, 3.21]. By the Sphere Theorem [116, Theorem 4.3], an irreducible 3-manifold is aspherical if and only if it is a 3-disk or has infinite fundamental group.

A properly-embedded orientable connected surface in a 3-manifold is *incompressible* if it is not a 2-sphere and the inclusion induces an injection on the fundamental groups. One says that ∂M is *incompressible in* M if and only if ∂M is empty or any component C of ∂M is incompressible in the sense above. An irreducible 3-manifold is *Haken* if it contains an embedded orientable incompressible surface. If M is irreducible, and in addition $H_1(M)$ is infinite, which is implied if ∂M contains a surface other than S^2, then M is Haken [116, Lemma 6.6 and 6.7]. (With our definitions, any properly embedded 2-disk is incompressible, and the 3-disk is Haken.)

We call a manifold *hyperbolic* if its interior admits a complete Riemannian metric whose sectional curvature is constant -1. Provided that M has no boundary, this is equivalent to the statement that the universal covering with the lifted Riemannian metric is isometrically isomorphic to the hyperbolic space of the same dimension as M. We use the definition of a Seifert 3-manifold of [231, p. 429]. If $\pi_1(M)$ is infinite, M is Seifert if and only if some finite covering of M is the total space of an S^1-principal bundle over a compact 2-dimensional manifold. The work of Casson and Gabai shows that an irreducible 3-manifold with infinite fundamental group π is Seifert if and only if π contains a normal infinite cyclic subgroup [98, Corollary 2, p. 395].

Next we mention what is known about Thurston's Geometrization Conjecture for irreducible 3-manifolds with infinite fundamental groups. Johannson [127] and Jaco and Shalen [125] have shown that, given an irreducible 3-manifold M with incompressible boundary, there is a finite family of disjoint, pairwise-nonisotopic incompressible tori in M which are not isotopic to boundary components and which split M into pieces that are Seifert manifolds or are geometrically atoroidal, meaning that they admit no embedded incompressible torus (except possibly parallel to the boundary). A minimal family of such tori is unique up to isotopy, and we will say that it gives a *toral splitting* of M. We will say that the toral splitting is *geometric* if the geometrically atoroidal pieces which do not admit a Seifert structure are hyperbolic. *Thurston's Geometrization Conjecture* for irreducible 3-manifolds with infinite fundamental groups states that such manifolds have geometric toral splittings. For completeness we mention that Thurston's Geometrization Conjecture says, for a closed 3-manifold with finite fundamental group, that its universal covering is homeomorphic to S^3, the fundamental group of M is a subgroup of $SO(4)$ and the action of it on the universal covering is conjugated by a homeomorphism to the restriction of the obvious $SO(4)$-action on S^3. This implies, in particular, the *Poincaré Conjecture* that any homotopy 3-sphere is homeomorphic to S^3.

Suppose that M is Haken. The pieces in its toral splitting are certainly Haken. Let N be a geometrically atoroidal piece. The *Torus Theorem* says that N is a special Seifert manifold or is homotopically atoroidal, i.e., any subgroup of $\pi_1(N)$ which is isomorphic to $\mathbb{Z} \times \mathbb{Z}$ is conjugate to the fundamental group of a boundary component. Thurston has shown that a homotopically atoroidal Haken manifold is a twisted I-bundle over the Klein bottle (which is Seifert) or is hyperbolic.

Thus the case in which Thurston's Geometrization Conjecture for an irreducible 3-manifold M with infinite fundamental group is still open is when M is a closed non-Haken irreducible 3-manifold with infinite fundamental group which is not Seifert. The conjecture states that such a manifold is hyperbolic.

We want to compute the L^2-Betti numbers of the universal covering of a 3-manifold under the assumption that Thurston's Geometrization Conjecture holds for the pieces in the prime decomposition with infinite fundamental group. First we deal with the case where the fundamental group of M is finite. In this case the L^2-Betti numbers are the ordinary Betti numbers of $\widetilde{M}$ divided by the order $|\pi|$ of the fundamental group π of M. If M is closed, $\widetilde{M}$ is homotopy equivalent to S^3, and hence $b_0^{(2)}(\widetilde{M}) = b_3^{(2)}(\widetilde{M}) = |\pi|^{-1}$ and $b_p^{(2)}(\widetilde{M}) = 0$ for $p \neq 0, 3$. Suppose that ∂M is non-trivial. Then M is a connected sum of homotopy spheres and k disks for some positive integer k, and hence $b_0^{(2)}(\widetilde{M}) = |\pi|^{-1}$, $b_2^{(2)}(\widetilde{M}) = (k-1) \cdot |\pi|^{-1}$ and $b_p^{(2)}(\widetilde{M}) = 0$ for $p \neq 0, 2$. Hence we only have to treat the case where $\pi_1(M)$ is infinite.

Let us say that a prime 3-manifold is *exceptional* if it is closed and no finite covering of it is homotopy equivalent to a Haken, Seifert or hyperbolic 3-manifold. No exceptional prime 3-manifolds are known, and Thurston's Geometrization Conjecture and *Waldhausen's Conjecture* that any 3-manifold is finitely covered by a Haken manifold imply that there are none. Notice that any exceptional manifold has infinite fundamental group.

THEOREM 3.3 (Lott and Lück [148, Theorem 0.1]). *Let M be the connected sum $M_1 \sharp \cdots \sharp M_r$ of (compact connected orientable) non-exceptional prime 3-manifolds M_j. Assume that $\pi_1(M)$ is infinite. Then the L^2-Betti numbers of the universal covering $\widetilde{M}$ are given by*

$$\begin{aligned}
b_0^{(2)}(\widetilde{M}) &= 0;\\
b_1^{(2)}(\widetilde{M}) &= (r-1) - \sum_{j=1}^{r} \frac{1}{|\pi_1(M_j)|} - \chi(M) + \left|\{C \in \pi_0(\partial M) \mid C \cong S^2\}\right|;\\
b_2^{(2)}(\widetilde{M}) &= (r-1) - \sum_{j=1}^{r} \frac{1}{|\pi_1(M_j)|} + \left|\{C \in \pi_0(\partial M) \mid C \cong S^2\}\right|;\\
b_3^{(2)}(\widetilde{M}) &= 0.
\end{aligned}$$

PROOF. We give a sketch of the strategy of proof. Since the fundamental group is infinite, we get $b_0^{(2)}(\widetilde{M}) = 0$ from Theorem 1.7.9. If M is closed, we get $b_3^{(2)}(\widetilde{M}) = 0$ because of Poincaré duality 1.7.3. If M has boundary, it is homotopy equivalent to a 2-dimensional CW-complex and hence $b_3^{(2)}(\widetilde{M}) = 0$. It remains to compute the second L^2-Betti number, because the first one is then determined by the Euler–Poincaré formula 1.7.2.

It is not hard to find a general formula for the L^2-Betti numbers of a connected sum in terms of the summands. With this formula we reduce the claim to prime 3-manifolds. Since a prime 3-manifold is either irreducible or $S^1 \times S^2$, the claim is reduced to the irreducible case. If the boundary is compressible, we use the Loop Theorem [116, Theorem 4.2, p. 39] to reduce the claim to the incompressible case. By doubling M we can reduce the claim further to the case of an irreducible 3-manifold with infinite fundamental group and incompressible torus boundary. By the toral splitting and the assumptions about Thurston's Geometrization Conjecture we can reduce further to the claim that the L^2-Betti numbers vanish if M is Seifert with infinite fundamental group or is hyperbolic with incompressible torus boundary. All these steps use the weakly exact Mayer–Vietoris sequence for L^2-(co)-homology of Cheeger and Gromov [52, Theorem 2.1, p. 10]. In the Seifert case we can assume by the multiplicative property (see Theorem 1.7.7) that M is an S^1-principal bundle over a 2-dimensional manifold. Then we use induction over the cells of the base space and the fact that the L^2-Betti numbers of S^1 vanish. The hyperbolic case is reduced to the known closed case by a careful analysis at the boundary using the fact that such a hyperbolic manifold with incompressible torus boundary has finite volume. □

Let $\chi_{\mathrm{virt}}(\pi_1(M))$ be the rational-valued group Euler characteristic of the group $\pi_1(M)$ in the sense of [35, IX.7], [246]. Then the conclusion in Theorem 3.3 is equivalent to

$$\begin{aligned} b_1^{(2)}(\widetilde{M}) &= -\chi_{\mathrm{virt}}(\pi_1(M)); \\ b_2^{(2)}(\widetilde{M}) &= \chi(M) - \chi_{\mathrm{virt}}(\pi_1(M)). \end{aligned}$$

This is proven in [148, pp. 53–54].

Notice that Theorem 3.3 proves Conjecture 2.6 and the third of the three equivalent assertions in Lemma 2.2 for compact 3-manifolds, provided that Thurston's Geometrization Conjecture or Waldhausen's Conjecture is true. Notice that this does *not* imply Conjecture 2.1 for Γ the fundamental group of a compact 3-manifold satisfying Thurston's or Waldhausen's Conjecture.

4. Aspherical manifolds and amenability

This section is devoted to a result of Cheeger and Gromov about the vanishing of the L^2-Betti numbers of the universal covering of an aspherical CW-complex whose fundamental group contains a non-trivial normal amenable subgroup.

Let $l^\infty(\Gamma, \mathbb{R})$ be the space of bounded functions from Γ to $\mathbb{R}$ with the supremum norm. Denote by 1 the constant function with value 1. A group Γ is called *amenable* if there is a Γ-invariant linear operator $\mu : l^\infty(\Gamma, \mathbb{R}) \to \mathbb{R}$ with $\mu(1) = 1$ which satisfies

$$\inf\{f(\gamma) \mid \gamma \in \Gamma\} \leqslant \mu(f) \leqslant \sup\{f(\gamma) \mid \gamma \in \Gamma\} \quad \text{for } f \in l^\infty(\Gamma).$$

The last condition is equivalent to the condition that μ is bounded and $\mu(f) \geqslant 0$ if $f(\gamma) \geqslant 0$ for all $\gamma \in \Gamma$.

We give an overview of some basic properties of this notion. The class of amenable groups satisfies the conditions appearing in the definition of elementary amenable groups in Section 2, namely, it contains all finite and all abelian groups, and is closed under taking subgroups, forming factor groups, group extensions, and upwards directed unions [17, Proposition F.6.11, p. 309]. Hence any elementary amenable group is amenable. Recently Grigorchuk has constructed a finitely presented group which is amenable but not elementary amenable. Any group containing the free group on two letters $\mathbb{Z} * \mathbb{Z}$ as subgroup is not amenable [17, Proposition F.6.12, p. 310]. There are finitely generated, but not finitely presented groups, which are not amenable but do not contain $\mathbb{Z} * \mathbb{Z}$ [200]. However, no non-amenable finitely presented group is known which does not contain $\mathbb{Z} * \mathbb{Z}$. A useful geometric characterization of amenable groups is given by the Fölner criterion [17, Theorem F.6.8, p. 308] which says that a finitely presented group Γ is amenable if and only if for any positive integer n, any connected closed Riemannian manifold M with fundamental group $\pi_1(M) = \Gamma$, and $\varepsilon > 0$, there is a domain $\Omega \subset \widetilde{M}$ with $(n-1)$-measurable boundary such that the $(n-1)$-measure of $\partial\Omega$ does not exceed ε times the measure of Ω. Such a domain can be constructed by an appropriate finite union of translations of a fundamental domain if Γ is amenable. The fundamental group of a closed connected manifold is not amenable if M admits a Riemannian metric of non-positive curvature which is not zero everywhere [9]. A group is amenable if and only if all its finitely generated subgroups are amenable [17, Proposition F.6.11, p. 309]. Any finitely generated group which is not amenable has exponential growth [17, Proposition F.6.24, p. 318]. A group Γ is amenable if and only if the canonical map from the full C^*-algebra of Γ to the reduced C^*-algebra of Γ is an isomorphism [207, Theorem 7.3.9, p. 243]. A group Γ is amenable if and only if the reduced C^*-algebra of Γ is nuclear [139]. For more information about amenable groups we refer to [206].

The next result gives a positive answer to the Singer Conjecture 2.6 for special fundamental groups.

THEOREM 4.1 (Cheeger and Gromov [54]). *If X is an aspherical connected CW-complex of finite type such that its fundamental group contains a non-trivial normal amenable subgroup, then we get for the universal covering $\widetilde{X}$*

$$b_p^{(2)}(\widetilde{X}) = 0 \quad \text{for } p \geqslant 0.$$

In this section we will explain only one of the decisive steps in the proof of Theorem 4.1 in order to illustrate the meaning of the condition about amenability. We will complete the proof in Section 10.

Let X be a finite CW-complex with regular covering $\overline{X} \to X$ with group of deck transformations Γ. Let $C^*(\overline{X})$ be the dual cochain complex with complex coefficients $\hom_{\mathbb{Z}}(C_*(\overline{X}), \mathbb{C})$ of the cellular chain complex $C_*(\overline{X})$. An element in $C^p(\overline{X})$ is a function $u : S_p(\overline{X}) \to \mathbb{C}$ from the set of p-cells of $\overline{X}$ to the complex numbers. We call u *square-summable* if $\sum_{\bar{e} \in S_p(\overline{X})} |u(\bar{e})|^2$ is finite. The square-summable elements form a subcomplex $C^*_{(2)}(\overline{X}) \subset C^*(\overline{X})$. We equip the chain modules of $C^*_{(2)}(\overline{X})$ with the obvious Hilbert space structure. This definition agrees with Definition 1.5. Let $\mathcal{H}^p_{(2)}(\overline{X})$ be the Hilbert submodule of $C^p_{(2)}(\overline{X})$ consisting of those elements u which satisfy $c^p_{(2)}(u) = 0$ and $(c^{p-1}_{(2)})^*(u) = 0$. There are obvious maps

$$i^p : \mathcal{H}^p_{(2)}(\overline{X}) \to H^p_{(2)}(C^*_{(2)}(\overline{X})) = H^p_{(2)}(\overline{X}); \tag{4.2}$$

$$j^p : \mathcal{H}^p_{(2)}(\overline{X}) \to H^p(C^*(\overline{X})) = H^p(\overline{X}; \mathbb{C}). \tag{4.3}$$

The first map (4.2) is always an isomorphism by an elementary argument.

LEMMA 4.4 (Cheeger and Gromov [54, Lemma 3.1, p. 203]). *The map j^p of* (4.3) *is injective, provided that Γ is amenable.*

PROOF. Let $\overline{D} \subset \overline{X}$ be a fundamental domain for the Γ-action on X. Such a $\overline{D}$ is constructed by choosing for each closed p-cell $e \in S_p(X)$ one lift $\overline{e} \in S_p(\overline{X})$ and taking the union of these lifts for all $p \geqslant 0$. Because Γ is amenable, one can find a sequence of subcomplexes $\overline{X_j} \subset \overline{X}$ and natural numbers n_j and ∂n_j such that $\overline{X_j}$ is the union of n_j translates with distinct elements in $\gamma_j^k \in \Gamma$ for $k = 1, 2, \ldots, n_j$ of $\overline{D}$, ∂n_j is the numbers of cells of $\overline{D}$ which meet the boundary (in the sense of point set topology) of $\overline{X_j}$ and

$$\lim_{j \to \infty} \frac{\partial n_j}{n_j} = 0. \tag{4.5}$$

Let $K^p \subset \mathcal{H}^p_{(2)}(\overline{X})$ be the kernel of j^p. We have to show that K^p is trivial. Let pr_{K^p} (respectively pr^p_j) be the orthogonal projection $C^p_{(2)}(\overline{X}) \to C^p_{(2)}(\overline{X})$ onto K^p (respectively onto the complex subspace of square-summable cochains u which are supported on $\overline{X_j}$, i.e., $u(e) = 0$ unless e belongs to $\overline{X_j}$). Let $\chi_{\overline{e}}$ denote the characteristic function $S_p(\overline{X}) \to \mathbb{C}$ for a given $\overline{e} \in S_p(\overline{X})$. We get

$$\begin{aligned}
\dim_{\mathcal{N}(\Gamma)}(K^p) &= \mathrm{tr}_{\mathcal{N}(\Gamma)}(\mathrm{pr}_{K^p}) = \sum_{e \in S_p(X)} \langle \mathrm{pr}_{K^p}(\chi_{\overline{e}}), \chi_{\overline{e}} \rangle \\
&= \sum_{\overline{e} \subset S_p(\overline{D})} \langle \mathrm{pr}_{K^p}(\chi_{\overline{e}}), \chi_{\overline{e}} \rangle = \frac{1}{n_j} \cdot \sum_{\overline{e} \subset S_p(\overline{D})} n_j \cdot \langle \mathrm{pr}_{K^p}(\chi_{\overline{e}}), \chi_{\overline{e}} \rangle \\
&= \frac{1}{n_j} \cdot \sum_{\overline{e} \in S_p(\overline{D})} \sum_{k=1}^{n_j} \langle \mathrm{pr}_{K^p}(\chi_{\gamma_j^k \overline{e}}), \chi_{\gamma_j^k \overline{e}} \rangle \\
&= \frac{1}{n_j} \cdot \sum_{\overline{e} \in S_p(\overline{X_j})} \langle \mathrm{pr}_{K^p}(\chi_{\overline{e}}), \chi_{\overline{e}} \rangle = \frac{1}{n_j} \cdot \mathrm{tr}_{\mathbb{C}}(\mathrm{pr}^p_j \circ \mathrm{pr}_{K^p}).
\end{aligned}$$

Since pr^p_j and pr_{K^p} have norm 1, we conclude

$$\dim_{\mathcal{N}(\Gamma)}(K^p) \leqslant \frac{1}{n_j} \cdot \dim_{\mathbb{C}}(\mathrm{im}(\mathrm{pr}^p_j \circ \mathrm{pr}_{K^p})). \tag{4.6}$$

We fix the orthogonal decomposition

$$u = u_i + u_\partial + u_e \in C^p_{(2)}(\overline{X}) \subset \mathrm{map}(S_p(\overline{X}), \mathbb{C}),$$

where u_i is only supported on (closed) cells in the interior of $\overline{X_j}$, u_∂ is supported on cells which meet the boundary of $\overline{X_j}$ and u_e is supported on cells which do not meet $\overline{X_j}$ (and hence do not meet the boundary of $\overline{X_j}$).

Now consider an element $u \in K^p$ such that $u_\partial = 0$. Choose $v \in C^{p-1}(\overline{X})$ such that $c^{p-1}: C^{p-1}(\overline{X}) \to C^p(\overline{X})$ maps v to u regarded as an element in $C^p(\overline{X})$. Notice that pr_j^p extends to a map $\mathrm{pr}_j^p : C^p(\overline{X}) \to C^p(\overline{X})$, namely $\mathrm{pr}_j^p(y)$ for $y \in C^p(\overline{X})$ sends a cell $e \in S_p(\overline{X})$ to $y(e)$ if e belongs to $\overline{X_j}$ and to zero otherwise. From this description we conclude $\mathrm{pr}_j^p(u) = u_i$. The cochain $\mathrm{pr}_j^p \circ c^{p-1}(v)$ vanishes on cells which do not meet $\overline{X_j}$. The difference $c^{p-1} \circ \mathrm{pr}_j^{p-1}(v) - \mathrm{pr}_j^p \circ c^{p-1}(v)$ is supported on cells which meet the boundary of $\overline{X_j}$ because the boundary of a cell which does not belong to $\overline{X_j}$ cannot lie in the interior of $\overline{X_j}$. Hence we get

$$\begin{aligned} 0 &\leqslant \langle \mathrm{pr}_j^p(u), \mathrm{pr}_j^p(u)\rangle = \langle \mathrm{pr}_j^p \circ c^{p-1}(v), u_i\rangle \\ &= \langle c^{p-1} \circ \mathrm{pr}_j^{p-1}(v), u_i\rangle = \langle c^{p-1} \circ \mathrm{pr}_j^{p-1}(v), u\rangle \\ &= \langle \mathrm{pr}_j^{p-1}(v), (c^{p-1})^*(u)\rangle = \langle \mathrm{pr}_j^{p-1}(v), 0\rangle = 0. \end{aligned}$$

This implies that pr_j vanishes on $K^p \cap \{u \in C_{(2)}^p(\overline{X}) \mid u_\partial = 0\}$. Since

$$\begin{aligned} &\dim_{\mathbb{C}}\bigl(K^p/\bigl(K^p \cap \{u \in C_{(2)}^p(\overline{X}) \mid u_\partial = 0\}\bigr)\bigr) \\ &\quad \leqslant \dim_{\mathbb{C}}\bigl(C_{(2)}^p(\overline{X})/\{u \in C_{(2)}^p(\overline{X}) \mid u_\partial = 0\}\bigr) \\ &\quad = \partial n_j, \end{aligned}$$

we conclude

$$\dim_{\mathbb{C}}\bigl(\mathrm{im}\bigl(\mathrm{pr}_j^p \circ \mathrm{pr}_{K^p}\bigr)\bigr) \leqslant \partial n_j. \tag{4.7}$$

We get from (4.6) and (4.7)

$$\dim_{\mathcal{N}(\Gamma)}\bigl(K^p\bigr) \leqslant \frac{\partial n_j}{n_j}.$$

Since the limit of the right hand side is zero by (4.5), we get

$$\dim_{\mathcal{N}(\Gamma)}\bigl(K^p\bigr) = 0.$$

This finishes the proof of Lemma 4.4. □

Notice that Lemma 4.4 proves Theorem 4.1 in the special case where the fundamental group of X is itself amenable by the following argument. In order to show that $b_p^{(2)}(\overline{X})$ vanishes, we can pass to the $(d+1)$-skeleton $\overline{Y} \subset \overline{X}$ and prove that $b_p^{(2)}(\overline{Y})$ vanishes. Since $\overline{X}$ is aspherical, we conclude

$$H^p(\overline{Y}; \mathbb{C}) = H^p(\overline{X}; \mathbb{C}) = 0.$$

Now Lemma 4.4 implies

$$H^p_{(2)}(\overline{Y}) = 0; \qquad b^p_{(2)}(\overline{Y}) = 0.$$

If Γ contains a normal amenable infinite group Δ such that $B\Delta$ is of finite type, then Theorem 4.1 follows by the L^2-version of the Leray–Serre spectral sequence (see, for instance, [239]) applied to the fibration $B\Delta \to B\Gamma \to B\Gamma/\Delta$. We will give the proof in the general case, where no assumptions about $B\Delta$ are made, in Section 10.

5. Approximating L^2-Betti numbers by ordinary Betti numbers

In this section we get the L^2-Betti numbers of a regular covering $\overline{X} \to X$ of a CW-complex of finite type as the limit of the normalized ordinary Betti numbers of a tower of finite coverings $X_m \to X$ which converges in some sense to $\overline{X}$.

Let $\overline{X} \to X$ be a regular covering of a CW-complex of finite type with group of deck transformations Γ. Suppose that Γ is *residually finite*, i.e., for each element $\gamma \in \Gamma$ with $\gamma \neq 1$ there is a homomorphism $\phi : \Gamma \to G$ to a finite group with $\phi(\gamma) \neq 1$. We will assume that Γ is countable. Under this assumption Γ is residually finite if and only if there is a nested sequence of normal in Γ subgroups $\cdots \subset \Gamma_{m+1} \subset \Gamma_m \subset \cdots \subset \Gamma_0 = \Gamma$ such that the index $[\Gamma : \Gamma_m]$ is finite for all $m \geqslant 0$ and the intersection $\bigcap_{m\geqslant 0} \Gamma_m$ is the trivial group. We give some information about residually finite groups at the end of this section. Consider any such sequence $(\Gamma_m)_{m\geqslant 0}$. Let $p_m : X_m = \Gamma_m \backslash \overline{X} \to X$ be the covering of X associated with $\Gamma_m \subset \Gamma$. Notice that this is a finite regular covering of X, and hence X_m is again of finite type. Denote by $b_p(X_m)$ the (ordinary) pth Betti number of X_m.

THEOREM 5.1 (Lück [156]). *Under the conditions above we get for all $p \geqslant 0$*

$$\lim_{m\to\infty} \frac{b_p(X_m)}{[\Gamma : \Gamma_m]} = b_p^{(2)}(\overline{X}).$$

The inequality $\limsup_{m\to\infty} \frac{b_p(X_m)}{[\Gamma:\Gamma_m]} \leqslant b_p^{(2)}(\overline{X})$ for X a closed manifold is discussed by Gromov [109, 0.5.F, 8.A] and is essentially due to Kazhdan [132]. The paper of Donnelly [77] deals with the operator $f(\Delta_0)$ acting on 0-forms for a function $f \in C_0^\infty(\mathbb{R}^+)$. Theorem 5.1 is proven by Yeung [254] in the special case of a closed Kähler manifold with negative sectional curvature and by DeGeorge–Wallach [102,103] in the special case of a closed locally symmetric space of non-compact type. In the last case all the L^2-Betti numbers vanish, so one gets

$$\lim_{m\to\infty} \frac{b_p(X_m)}{[\Gamma : \Gamma_m]} = 0.$$

This leads to the question of how fast this sequence goes to zero; in other words, one wants to know the largest ε for which

$$\lim_{m\to\infty} \frac{b_p(X_m)}{[\Gamma : \Gamma_m]^{1-\varepsilon}} = 0$$

is true. Such questions are treated in [227,251,252].

The general case of a CW-complex X of finite type is proven in [156]. There actually the entire spectral density function of the combinatorial Laplace operator on the cellular L^2-chain complex of $\overline{X}$ is approximated by the spectral density function for the combinatorial Laplace operator on the cellular chain complexes of the various X_m. The spectral density function for X_m simply encodes the eigenvalues and their multiplicities of the Laplace operator because X_m is compact, whereas the one for $\overline{X}$ is more complicated as, in general, $\overline{X}$ is not compact and the spectrum is not discrete. In some sense we try to approximate continuous information by discrete data. The philosophy is that the tower of finite coverings X_m converges to $\overline{X}$.

The inequality $\limsup_{m\to\infty} \frac{b_p(X_m)}{[\Gamma:\Gamma_m]} \leqslant b_p^{(2)}(\overline{X})$ is the easier part of the proof of Theorem 5.1. The main trick in the proof of the other inequality is not to forget and to use essentially the fact that the combinatorial Laplace operator on the cellular chain complex already lives over the integral group ring. This allows the use of the obvious inequality $|n| \geqslant 1$ for an integer n different from zero. The results for the combinatorial Laplace operator on the cellular chain complexes carry over to the analytic Laplace operator acting on smooth p-forms on X provided that X is a compact smooth manifold. One can formulate Theorem 5.1 for any elliptic differential operator on a closed smooth Riemannian manifold X if one uses on each X_m and on $\overline{X}$ the lifted Riemannian metric and elliptic differential operator and substitutes the Betti numbers by the dimensions of the kernels. It seems not hard to prove the analogue of the inequality $\limsup_{m\to\infty} \frac{b_p(X_m)}{[\Gamma:\Gamma_m]} \leqslant b_p^{(2)}(\overline{X})$, but it is not known whether the equality holds. The question is whether the result is only true for the Laplacian because it has a cellular analogue which allows us to use the fact that everything already lives over the integral group ring. Again we see from this discussion as we have mentioned in Section 2 that the passage from the integral group ring of Γ to the reduced C^*-algebra, or even the von Neumann algebra of Γ, plays a fundamental role and any understanding of it seems to give new results.

Finally we collect some basic facts about residually finite groups in order to explain how restrictive the assumption is that Γ is residually finite. For more information we refer to the survey article of Magnus [169].

The free product of two residually finite groups is again residually finite [60, p. 27], [113]. A finitely generated residually finite group has a solvable word problem [183]. The automorphism group of a finitely generated residually finite group is residually finite [14]. A finitely generated residually finite group is Hopfian, i.e., any surjective endomorphism is an automorphism [168], [195, Corollary 41.44]. Let Γ be a finitely generated group possessing a faithful representation into $GL(n, F)$ for F a field. Then Γ is residually finite [168], [248, Theorem 4.2]. The fundamental group of a compact 3-manifold whose prime decomposition consists of manifolds which have finite fundamental groups, or are non-exceptional in the sense of Section 3 (i.e., which are finitely covered by a manifold which is homotopy equivalent to a Haken, Seifert or hyperbolic manifold), is residually finite [117, p. 380]. Let Γ be a finitely generated group. Let Γ^{rf} be the quotient of Γ by the normal subgroup which is the intersection of all normal subgroups of Γ of finite index. The group Γ^{rf} is residually finite and any finite-dimensional representation of Γ over a field factorizes over the canonical projection $\Gamma \to \Gamma^{rf}$.

The upshot of this discussion is the slogan that the fundamental group of a geometrically interesting closed manifold is very likely to be residually finite. However, there is an infinite group Γ with four generators and four relations which has no finite quotient except the trivial one and hence satisfies $\Gamma^{rf} = \{1\}$ [119].

Meanwhile Theorem 5.1 has been generalized and put into somewhat different context in [56,74,92,229].

6. L^2-Betti numbers and groups

In this section we explain some applications of L^2-Betti numbers to group theory. Given a group Γ, we define its L^2-Betti number $b_p^{(2)}(\Gamma)$ by $b_p^{(2)}(E\Gamma)$ where $E\Gamma \to B\Gamma$ is the universal principal Γ-bundle. This is only well-defined if the $(p+1)$-skeleton of $B\Gamma$ is finite. The definition for arbitrary groups will be given in Definition 10.9.

The next result of Cheeger and Gromov was proven in special cases by Gottlieb [104] and Rosset [219].

THEOREM 6.1 (Cheeger and Gromov [54, Corollary 0.6, p. 193]). *Let Γ be a group such that there exists a subgroup Γ' of finite index whose classifying space $B\Gamma'$ is a finite CW-complex. Let $\chi_{\mathrm{virt}}(\Gamma)$ be the rational-valued virtual Euler characteristic of the group Γ in the sense of* [35, IX.7], [246]. *Suppose that Γ contains an infinite normal amenable subgroup. Then*

$$\chi_{\mathrm{virt}}(\Gamma) = 0.$$

PROOF. By definition $\chi_{\mathrm{virt}}(\Gamma) = \frac{1}{[\Gamma:\Gamma']} \cdot \chi(B\Gamma')$. We derive from the Euler–Poincaré formula of Theorem 1.7.2

$$\chi_{\mathrm{virt}}(\Gamma) = \frac{1}{[\Gamma:\Gamma']} \cdot \sum_{p \geqslant 0} (-1)^p \cdot b_p^{(2)}(B\Gamma').$$

Now the claim follows from Theorem 4.1. □

We mention the following result

THEOREM 6.2 (Reich [213, Corollary 9.3]). *Let Γ be an infinite group which belongs to Linnell's class $\mathcal{C}$ introduced in Section 2 and has an upper bound on the orders of its finite subgroups. Suppose that there exists a subgroup Γ' of finite index whose classifying space $B\Gamma'$ is a finite CW-complex. Then*

$$\chi_{\mathrm{virt}}(\Gamma) \leqslant 0.$$

Next we mention the following observation about Thompson's group F. It consists of orientation preserving dyadic PL-automorphisms of $[0, 1]$, where dyadic means that all slopes are integral powers of 2 and the break points are contained in $\mathbb{Z}[1/2]$. It has the presentation

$$F = \langle x_0, x_1, x_2, \ldots \mid x_i^{-1} x_n x_i = x_{n+1} \text{ for } i < n \rangle.$$

This group has some very interesting properties. It is not elementary amenable and does not contain a subgroup which is free on two generators [33,45]. Hence it is a very interesting question whether F is amenable. Since BF is of finite type [36], the L^2-Betti numbers $b_p^{(2)}(F)$ are defined for all $p \geqslant 0$. We conclude from Theorem 4.1 of Cheeger and Gromov that a necessary condition for F to be amenable is that $b_p^{(2)}(F)$ vanishes for all $p \geqslant 0$. This motivates the following result.

THEOREM 6.3 (Lück [157, Theorem 0.8]). *All the L^2-Betti numbers $b_p^{(2)}(F)$ of Thompson's group F vanish.*

For the proof of Theorem 6.3 we need the next result.

Given a selfmap $f: F \to F$, its *mapping cylinder* M_f is obtained by gluing the bottom of the cylinder $F \times [0,1]$ to F by the identification $(x,0) = f(x)$. Its *mapping torus* T_f is obtained from the mapping cylinder by identifying the top and the bottom by the identity. If f is a homotopy equivalence, T_f is homotopy equivalent to the total space of a fibration over S^1 with fiber F. Conversely, the total space of such a fibration is homotopy equivalent to the mapping torus of the self homotopy equivalence of F given by the fiber transport with a generator of $\pi_1(S^1)$. The homotopy type of T_f depends only on the homotopy class of f. There is an obvious map from T_f to S^1 which induces an epimorphism $\mu: \pi_1(T_f) \to \mathbb{Z}$.

THEOREM 6.4 (Lück [155, Theorem 2.1, p. 207]). *Let F be a connected CW-complex of finite type and $f: F \to F$ be a selfmap. Let*

$$\mu: \pi_1(T_f) \xrightarrow{\phi} \Gamma \xrightarrow{\psi} \mathbb{Z}$$

be a factorization of μ into epimorphisms. Let $\overline{T_f} \to T_f$ be the regular covering of T_f with Γ as group of deck transformations which is associated to ϕ. Then

$$b_p^{(2)}(\overline{T_f}) = 0 \quad \textit{for } p \geqslant 0.$$

PROOF. For simplicity we give here only the proof in the case where f is a homotopy equivalence and ϕ is the identity on $\pi_1(T_f)$, i.e., $\overline{T_f}$ is the universal covering $\widetilde{T_f}$. Consider any positive integer d. Let Γ_d be the preimage of $d\mathbb{Z}$ under $\mu: \pi_1(T_f) \to \mathbb{Z}$. Let $\overline{\overline{T_f}}$ be the covering $\widetilde{T_f} \to \Gamma_d \backslash \widetilde{T_f}$ with Γ_d as group of deck transformations. We get from the multiplicative property of Theorem 1.7.7

$$b_p^{(2)}(\widetilde{T_f}) = \frac{1}{d} \cdot b_p^{(2)}(\overline{\overline{T_f}}).$$

One easily checks that $\Gamma_d \backslash \widetilde{T_f}$ is homotopy equivalent to T_{f^d}. On T_{f^d} there is a CW-complex structure whose number of p-cells is bounded by the number C which is the sum of the numbers of p-cells and the number of $(p-1)$-cells in F. Notice that C is independent of d. Since the pth chain module in the cellular L^2-chain complex of $\overline{\overline{T_{f^d}}}$ is $\bigoplus_{i=1}^{C} l^2(\Gamma_d)$, we conclude

$$b_p^{(2)}\big(\overline{\overline{T_{f^d}}}\big) = \dim_{\mathcal{N}(\Gamma_d)}\big(H_p^{(2)}\big(\overline{\overline{T_{f^d}}}\big)\big) \leqslant \dim_{\mathcal{N}(\Gamma_d)}\big(C_p^{(2)}\big(\overline{\overline{T_{f^d}}}\big)\big) = C.$$

Hence we have shown that for all $d \geqslant 1$

$$0 \leqslant b_p^{(2)}(\widetilde{T_f}) \leqslant \frac{C}{d}.$$

Taking the limit for $d \to \infty$ finishes the proof of Theorem 6.4. □

Next we give the proof of Theorem 6.3. There is a subgroup $F_1 \subset F$ together with a monomorphism $\Phi : F_1 \to F_1$ such that F_1 is isomorphic to F and F is the HNN-extension of F_1 with respect to Φ with one stable letter [36, Proposition 1.7, p. 370]. From the topological description of HNN-extensions [166, p. 180] we conclude that F is the fundamental group of the mapping torus $T_{B\Phi}$ of the map $B\Phi : BF_1 \to BF_1$ induced by Φ. The inclusion $BF_1 \to BF$ induces on the fundamental groups the inclusion of F_1 in F. The calculation in [155, p. 207] shows that the cellular $\mathbb{Z}F$-chain complex of the universal covering $\widetilde{T_{B\Phi}}$ of $T_{B\Phi}$ is the mapping cone of a certain $\mathbb{Z}F$-chain map from $\mathbb{Z}F \otimes_{\mathbb{Z}F_1} C(EF_1)$ to itself. Since $\mathbb{Z}F$ is free over $\mathbb{Z}F_1$, we conclude for $p \geqslant 1$

$$H_p\big(\mathbb{Z}F \otimes_{\mathbb{Z}F_1} C(EF_1)\big) = \mathbb{Z}F \otimes_{\mathbb{Z}F_1} H_p\big(C(EF_1)\big) = 0.$$

This implies $H_p(\widetilde{T_{B\Phi}}; \mathbb{Z}) = 0$ for $p \geqslant 2$. Hence $T_{B\Phi}$ is a model for BF. Now Theorem 6.3 follows from Theorem 6.4. □

Next we treat the notion of deficiency. Let Γ be a finitely presented group. Its *deficiency* is the maximum over all differences $g - r$, where g respectively r is the number of generators respectively relations of a presentation of Γ. One can show that the maximum does exist. Sometimes the deficiency of a group is what one would guess from an obvious presentation as in the following cases

group	presentation	deficiency
$*_{i=1}^{g}\mathbb{Z}$	$\langle s_1, \ldots, s_g \mid \emptyset \rangle$	g
$\mathbb{Z}/n,\ n \geqslant 2$	$\langle s \mid s^n = 1 \rangle$	0
$\mathbb{Z}/n \times \mathbb{Z}/n,\ n \geqslant 2$	$\langle s, t \mid s^n = t^n = [s,t] = 1 \rangle$	-1

On the other hand, the group $(\mathbb{Z}/2 \times \mathbb{Z}/2) * (\mathbb{Z}/3 \times \mathbb{Z}/3)$ has the obvious presentation

$$\langle s_0, t_0, s_1, t_1 \mid s_0^2 = t_0^2 = [s_0, t_0] = s_1^3 = t_1^3 = [s_1, t_1] = 1 \rangle$$

and one may think that its deficiency is -2. However, it turns out that its deficiency is -1. For instance, there is the following presentation, which looks on the first glance to be the presentation above with one relation missing

$$\langle s_0, t_0, s_1, t_1 \mid s_0^2 = 1, [s_0, t_0] = t_0^2, s_1^3 = 1, [s_1, t_1] = t_1^3, t_0^2 = t_1^3 \rangle.$$

The following calculation shows that, from the five relations appearing in the presentation above, the relation $t_0^2 = 1$ follows which shows that the presentation above indeed is a presentation of $(\mathbb{Z}/2 \times \mathbb{Z}/2) * (\mathbb{Z}/3 \times \mathbb{Z}/3)$.

We start with proving inductively for $k = 1, 2, \ldots$ the relation $s_i^k t_i s_i^{-k} = t_i^{r_i^k}$ for $i = 0, 1$ where $r_0 = 3$ and $r_1 = 4$. The beginning of the induction is obvious, the induction step follows from the calculation

$$s_i^{k+1} t_i s_i^{-(k+1)} = s_i s_i^k t_i s_i^{-k} s_i^{-1} = s_i t_i^{r_i^k} s_i^{-1} = \left(s_i t_i s_i^{-1}\right)^{r_i^k} = \left(t_i^{r_i}\right)^{r_i^k} = t_i^{r_i^{k+1}}.$$

This implies, for $k = 2$, $i = 0$ and $k = 3$, $i = 1$

$$t_0 = t_0^{3^2}; \qquad t_1 = t_1^{4^3}.$$

Since $t_0^2 = t_1^3$, we conclude

$$\left(t_0^2\right)^4 = 1; \qquad \left(t_0^2\right)^{21} = 1.$$

As 4 and 21 are prime, we get $t_0^2 = 1$ and the claim follows.

The example above is a special case of a family of examples described by Hog-Ancheloni, Lustig and Metzler [122]. The example shows that the deficiency is not additive under free products in general. However, we believe that this is true for torsionfree finitely presented groups. The example above plays a fundamental role in the counterexample, up to homotopy, of the Kneser Conjecture in dimension 4 [137].

The link between the deficiency and the L^2-Betti numbers of a group is the following elementary lemma.

LEMMA 6.5. (1) *Let Γ be a finitely presented group. Then*

$$\operatorname{def}(\Gamma) \leqslant 1 - b_0^{(2)}(\Gamma) + b_1^{(2)}(\Gamma) - b_2^{(2)}(\Gamma).$$

(2) *If M is a closed oriented 4-manifold, then we get for its signature*

$$\left|\operatorname{sign}(M)\right| \leqslant b_2^{(2)}\left(\widetilde{M}\right).$$

PROOF. (1) Given a presentation with r relations and g generators, let X be the corresponding connected 2-dimensional CW-complex with fundamental group isomorphic to Γ which has precisely one cell of dimension 0, g cells of dimension 1 and r cells of dimension 2. Since the classifying map $X \to B\Gamma$ is 2-connected, we conclude from Theorem 1.7

$$\begin{aligned} 1 - g + r &= \chi(X) = b_0^{(2)}\left(\widetilde{X}\right) - b_1^{(2)}\left(\widetilde{X}\right) + b_2^{(2)}\left(\widetilde{X}\right) \\ &\geqslant b_0^{(2)}(\Gamma) - b_1^{(2)}(\Gamma) + b_2^{(2)}(\Gamma). \end{aligned}$$

(2) According to the L^2-signature theorem [3, p. 71], the signature $\sigma(M)$ is the difference of the von Neumann dimensions of two complementary subspaces of the second L^2-cohomology $H^2_{(2)}(\widetilde{M})$. This implies

$$|\mathrm{sign}(M)| \leqslant \dim_{\mathcal{N}(\pi_1(M))}\left(H^2_{(2)}(\widetilde{M})\right) = b_2^{(2)}(\widetilde{M}). \qquad \square$$

Lemma 6.5 has an analogous formulation if one uses ordinary L^2-Betti numbers with coefficients in any field. The values of the deficiencies of the groups in the list above follow from Lemma 6.5. In particular, one sees that the deficiency is defined as a natural number, i.e., there is an upper bound on the possible values $g - r$ appearing in the definition of deficiency. One also rediscovers the well-known fact that the deficiency of a finite group is less than or equal to zero.

If Γ is a torsion-free one-relator group, the 2-dimensional CW-complex associated with the presentation is aspherical and hence $B\Gamma$ is 2-dimensional [166, Chapter III, §§9–11]. If Γ has a presentation with g generators and one (non-trivial) relation, its deficiency is $g - 1$. We conjecture that for a torsion-free group having a presentation with $g \geqslant 2$ generators and one non-trivial relation $b_2^{(2)}(\Gamma) = 0$ and $b_1^{(2)}(\Gamma) = \mathrm{def}(\Gamma) - 1 = g - 2$ holds. This would follow from Conjecture 2.1. Namely, the kernel of the second differential of the L^2-chain complex of $\widetilde{E\Gamma}$ is a submodule of $l^2(\Gamma)$ so that its dimension $b_2^{(2)}(\Gamma)$ is less or equal to the dimension of $l^2(\Gamma)$ which is 1. Since Γ is by assumption torsionfree, the dimension of the kernel is an integer by Conjecture 2.1. The second differential in the cellular $\mathbb{Z}\Gamma$-chain complex cannot be trivial because the relation in Γ is assumed to be non-trivial. Hence the dimension of kernel of the second differential of the L^2-chain complex of $E\Gamma$ is trivial. This shows $b_2^{(2)}(\Gamma) = 0$. The Euler–Poincaré formula of Theorem 1.7.2 and Theorem 1.7.9 imply $b_0^{(2)}(\Gamma) = 0$ and $b_1^{(2)}(\Gamma) = g - 2$. We get from Lemma 6.5 that $\mathrm{def}(\Gamma) = g - 1$.

The following is a direct consequence of [87, Theorem 2.5]. Let M be a compact 3-manifold with fundamental group Γ and prime decomposition $M = M_1 \sharp M_2 \sharp \cdots \sharp M_r$. Let $s(M)$ be the number of prime factors M_i with non-empty boundary and $t(M)$ be the number of prime factors which are S^2-bundles over S^1. Denote by $\chi(M)$ the Euler characteristic. Then

$$\begin{aligned} \mathrm{def}(\pi_1(M)) &= \dim_{\mathbb{Z}/2}\left(H_1(\pi; \mathbb{Z}/2)\right) - \dim_{\mathbb{Z}/2}\left(H_2(\pi; \mathbb{Z}/2)\right) \\ &= s(M) + t(M) - \chi(M). \end{aligned}$$

THEOREM 6.6 (Lück [157, Theorem 0.7]). *Let $1 \to \Delta \to \Gamma \to \pi \to 1$ be an exact sequence of groups such that Δ is finitely generated and infinite, Γ is finitely presented and $\mathbb{Z}$ is a subgroup of π. Then*

(1) $b_1^{(2)}(\Gamma) = 0$;

(2) $\mathrm{def}(\Gamma) \leqslant 1$;

(3) *Let M be a connected closed orientable 4-manifold with Γ as fundamental group. Then we get for its signature* $\mathrm{sign}(M)$ *and Euler characteristic* $\chi(M)$

$$|\mathrm{sign}(M)| \leqslant \chi(M).$$

PROOF. Assertion (1) follows from Theorem 6.7 below applied to $B\Delta \to B\Gamma \to B\pi$. The other two assertions follow from the first one by Lemma 6.5. □

The structure of groups of deficiency greater than or equal to 2 is examined in [15]. Theorem 6.6 generalizes results of [80,81,120,128,250]. The key ingredient in the proof of Theorem 6.6 is the next result.

THEOREM 6.7 (Lück [157, Theorem 7.1]). *Let $d \geqslant 0$ be an integer. Let $F \to E \xrightarrow{p} B$ be a fibration of spaces such that F (respectively E) has the homotopy type of a connected CW-complex with finite d-skeleton (respectively $(d+1)$-skeleton). Suppose that the image of $\pi_1(F) \to \pi_1(E)$ is infinite and $\mathbb{Z}$ is a subgroup of $\pi_1(B)$. Then*

$$b_1^{(2)}(\widetilde{E}) = 0.$$

PROOF. We give only a sketch of the proof. The idea is to consider the L^2-version of the Leray–Serre type spectral sequence associated to the fibration which is explained in [157] (see also [239]). It suffices to show that its E^2-term vanishes on the p-axis and on the q-axis. The vanishing on the p-axis is essentially a consequence of Theorem 1.7.9 and the assumption that the image of $\pi_1(F) \to \pi_1(E)$ is infinite. Since $\mathbb{Z}$ is a subgroup of $\pi_1(B)$, there is a map $f : S^1 \to B$ inducing an injection on the fundamental groups. Let $p_0 : E_0 \to S^1$ be the pull back of $p : E \to B$ with f. A spectral sequence comparison argument shows that it suffices to prove the vanishing of the E^2-term of p_0 on the q-axis. Since S^1 is 1-dimensional, the E^2-term is the E^∞-term, so that this is equivalent to the vanishing of the L^2-homology of $\widetilde{E_0}$. This follows from Theorem 6.4. □

More information about groups with vanishing first L^2-Betti number can be found in [16]. See also [121,147].

7. Kähler hyperbolic manifolds

In this section we explain Gromov's notion of Kähler-hyperbolic manifolds and his computations of the L^2-Betti numbers of the universal coverings of closed Kähler hyperbolic manifolds. In particular, we prove the Hopf Conjecture 2.7 for closed Kähler manifolds. In this section all manifolds have no boundary and come with a complex structure.

We recall some basic facts about Kähler manifolds which are standard in the compact case and extend to the not necessarily compact but complete situation. More details can be found for instance in [249, Chapter V]. Let M be a (complex) manifold. Let h be a Hermitian metric on M. In particular, we have for each $x \in M$ a Hermitian form $h_x : T_xM \times T_xM \to \mathbb{C}$. This induces a Riemannian metric g on M and a 2-form called *fundamental 2-form* ω defined on M by

$$g_x = \Re(h_x) : T_xM \times T_xM \to \mathbb{R};$$
$$\omega_x = -\frac{1}{2} \cdot \Im(h_x) : T_xM \times T_xM \to \mathbb{R}.$$

DEFINITION 7.1. A *Kähler manifold M* is a complex manifold M with Hermitian metric h such that (M, g) is complete and ω is closed, i.e., $d\omega = 0$. In this context ω is called the *Kähler form*.

Let M be a Kähler manifold of complex dimension $m = \dim_{\mathbb{C}}(M)$ and real dimension $n = \dim_{\mathbb{R}}(M) = 2m$. Next we deal with its Hodge theory. We introduce the following notations and identifications:

$$T_xM \otimes_{\mathbb{R}} \mathbb{C} = \operatorname{res}_{\mathbb{R}} T_xM \otimes_{\mathbb{R}} \mathbb{C};$$

$$\operatorname{Alt}^p(T_xM \otimes_{\mathbb{R}} \mathbb{C})$$
$$= \{\text{alternating } p\text{-forms over the complex vector space } T_xM \otimes_{\mathbb{R}} \mathbb{C}\};$$

$$\operatorname{Alt}^p_{\mathbb{R}}(\operatorname{res}_{\mathbb{R}} T_xM, \operatorname{res}_{\mathbb{R}} \mathbb{C})$$
$$= \{\text{real alternating } p\text{-forms on } \operatorname{res}_{\mathbb{R}} T_xM \text{ with values in } \operatorname{res}_{\mathbb{R}} \mathbb{C}\};$$

$$\operatorname{Alt}^p(T_xM \otimes_{\mathbb{R}} \mathbb{C}) = \operatorname{Alt}^p_{\mathbb{R}}(\operatorname{res}_{\mathbb{R}} T_xM, \operatorname{res}_{\mathbb{R}} \mathbb{C});$$

$$\Omega^p(M) = C^\infty\big(\operatorname{Alt}^p(TM \otimes_{\mathbb{R}} \mathbb{C})\big)$$
$$= \{\text{smooth } p\text{-forms on the smooth manifold } M \text{ with values in } \operatorname{res}_{\mathbb{R}} \mathbb{C}\}.$$

Let $J_x : T_xM \to T_xM$ be multiplication with i and $(J_x \otimes_{\mathbb{R}} \mathrm{id}) : T_xM \otimes_{\mathbb{R}} \mathbb{C} \to T_xM \otimes_{\mathbb{R}} \mathbb{C}$ be the induced map. Let $(T_xM \otimes_{\mathbb{R}} \mathbb{C})'$ respectively $(T_xM \otimes_{\mathbb{R}} \mathbb{C})''$ be the eigenspace of $(J_x \otimes_{\mathbb{R}} \mathrm{id})$ for the eigenvalue i (respectively $-i$). We obtain identifications (respectively decompositions):

$$T_xM \otimes_{\mathbb{R}} \mathbb{C} = (T_xM \otimes_{\mathbb{R}} \mathbb{C})' \oplus (T_xM \otimes_{\mathbb{R}} \mathbb{C})'';$$
$$T_xM = (T_xM \otimes_{\mathbb{R}} \mathbb{C})';$$
$$\operatorname{Alt}^{p,q}(T_xM \otimes_{\mathbb{R}} \mathbb{C}) = \operatorname{Alt}^p\big((T_xM \otimes_{\mathbb{R}} \mathbb{C})'\big) \otimes_{\mathbb{C}} \operatorname{Alt}^q\big((T_xM \otimes_{\mathbb{R}} \mathbb{C})''\big);$$
$$\operatorname{Alt}^r(T_xM \otimes_{\mathbb{R}} \mathbb{C}) = \bigoplus_{p+q=r} \operatorname{Alt}^{p,q}(T_xM \otimes_{\mathbb{R}} \mathbb{C});$$
$$\Omega^{p,q}(M) = C^\infty\big(\operatorname{Alt}^{p,q}(TM \otimes_{\mathbb{R}} \mathbb{C})\big);$$
$$\Omega^r(M) = \bigoplus_{p+q=r} \Omega^{p,q}(M).$$

We denote the map induced by complex conjugation by

$$\bar{\ } : \Omega^{p,q}(M) \to \Omega^{p,q}(M)$$

and the *Hodge star operator* by

$$* : \Omega^{p,q}(M) \to \Omega^{m-p,m-q}(M).$$

Let $L^2\Omega^p(M)$ be the *Hilbert space of square-integrable p-forms on M* with respect to the inner product

$$\langle \omega, \eta \rangle = \omega \wedge *\overline{\eta} = \int \langle \omega, \eta \rangle_x \, \mathrm{d}\,\mathrm{vol}.$$

Let $d : \Omega^r(M) \to \Omega^{r+1}(M)$ be the *exterior differential*. Its *adjoint*

$$d^* = (-1)^{(n+1)r+1} \overline{*} d \overline{*} : \Omega^{r+1}(M) \to \Omega^r(M)$$

satisfies $\langle d\omega, \eta \rangle = \langle \omega, d^*\eta \rangle$. Define

$$\begin{aligned} &\partial : \Omega^{p,q}(M) \to \Omega^{p+1,q}(M); \\ &\overline{\partial} : \Omega^{p,q}(M) \to \Omega^{p,q+1}(M) \end{aligned}$$

as the composition

$$\Omega^{p,q}(M) \hookrightarrow \Omega^r(M) \xrightarrow{d} \Omega^{r+1}(M) \xrightarrow{\mathrm{pr}} \begin{cases} \Omega^{p+1,q}(M), \\ \Omega^{p,q+1}(M), \end{cases}$$

Define *Laplace operators*

$$\begin{aligned} \Delta &= dd^* + d^*d : \Omega^r(M) \to \Omega^r(M); \\ \Box &= \partial\partial^* + \partial^*\partial : \Omega^{p,q}(M) \to \Omega^{p,q}(M); \\ \overline{\Box} &= \overline{\partial}\overline{\partial}^* + \overline{\partial}^*\overline{\partial} : \Omega^{p,q}(M) \to \Omega^{p,q}(M). \end{aligned}$$

These operators are related as follows.

LEMMA 7.2. *If M is a Kähler manifold, then*

$$\begin{aligned} &\partial \circ \partial = 0; && d = \partial + \overline{\partial}; \\ &\overline{\partial} \circ \overline{\partial} = 0; && \Delta = 2 \cdot \Box = 2 \cdot \overline{\Box}. \end{aligned}$$

DEFINITION 7.3. Define the *space of harmonic L^2-forms* by

$$\begin{aligned} \mathcal{H}_{(2)}^{p,q}(M) &= \left\{ \omega \in \Omega^{p,q}(M) \,\middle|\, \overline{\Box}(\omega) = 0, \int \omega \wedge *\omega < \infty \right\}; \\ \mathcal{H}_{(2)}^{r}(M) &= \left\{ \omega \in \Omega^{r}(M) \,\middle|\, \Delta(\omega) = 0, \int \omega \wedge *\omega < \infty \right\}. \end{aligned}$$

THEOREM 7.4 (L^2-Hodge–deRham decomposition in the Kaehler case). *If M is Kähler, then*

$$\begin{aligned} L^2\Omega^r(M) &= \mathcal{H}_{(2)}^r(M) \oplus \overline{d\big(L^2\Omega^{r-1}(M)\big)} \oplus \overline{d^*\big(L^2\Omega^{r+1}(M)\big)}; \\ \mathcal{H}_{(2)}^r(M) &= \bigoplus_{p+q=r} \mathcal{H}_{(2)}^{p,q}(M). \end{aligned}$$

THEOREM 7.5 (L^2-Lefschetz Theorem). *Let M be Kähler with Kähler form ω and real dimension $n = 2m$. Define*

$$L^k : \Omega^r(M) \to \Omega^{r+2k}(M), \qquad \phi \mapsto \phi \wedge \omega^k.$$

Then

(1) *L^k commutes with d, d^* and Δ;*

(2) *L^k induces bounded operators*

$$L^k : L^2\Omega^r(M) \to L^2\Omega^{r+2k}(M);$$
$$L^k : \mathcal{H}^r(M) \to \mathcal{H}^{r+2k}(M);$$

(3) *These operators are quasi-isometries, i.e., $C^{-1} \cdot \|\phi\| \leqslant \|L^k(\phi)\| \leqslant C \cdot \|\phi\|$ for appropriate $C > 0$, and in particular injective for $2r + 2k \leqslant n$ and surjective for $2r + 2k \geqslant n$.*

PROOF. We give a sketch of the proof. The Kähler condition implies that $\omega_x^m \neq 0$ for $x \in M$ and ω is parallel with respect to the Levi-Civita connection of (M, g). If $w : I \to M$ is a path from x to y and T_w is the induced isometric parallel transport, then $T_w^* \omega_y = \omega_x$ and the following diagram commutes

$$\begin{array}{ccc}
\mathrm{Alt}^r(T_xM \otimes_{\mathbb{R}} \mathbb{C}) & \xrightarrow{L_x^k} & \mathrm{Alt}^{r+2k}(T_xM \otimes_{\mathbb{R}} \mathbb{C}) \\
{\scriptstyle T_w^*}\downarrow & & \downarrow{\scriptstyle T_w^*} \\
\mathrm{Alt}^r(T_yM \otimes_{\mathbb{R}} \mathbb{C}) & \xrightarrow[L_y^k]{} & \mathrm{Alt}^{r+2k}(T_yM \otimes_{\mathbb{R}} \mathbb{C})
\end{array}$$

Linear algebra shows that $L_x^k : \mathrm{Alt}^r(T_xM \otimes_{\mathbb{R}} \mathbb{C}) \to \mathrm{Alt}^{r+2k}(T_xM \otimes_{\mathbb{R}} \mathbb{C})$ is injective for $2r + 2k \leqslant n$ and surjective for $2r + 2k \geqslant n$. Now everything follows except for the surjectivity statement for L^k. It suffices to prove surjectivity in the case $2r + 2k = n$ because of the factorization

$$L^{l+k} : \Omega^{r-2l}(M) \xrightarrow{L^l} \Omega^r(M) \xrightarrow{L^k} \Omega^{r+2k}(M).$$

If M is compact, the claim follows for $\mathcal{H}^r(M)$ as the Hodge star operator yields an isomorphism $\mathcal{H}^r(M) \cong \mathcal{H}^{r+2k}(M)$ and L^k is an injective map $\mathcal{H}^r(M) \to \mathcal{H}^{r+2k}(M)$ of finite dimensional complex vector spaces of the same dimension. In general, one argues as follows. Consider the adjoint of L_x^k

$$K_x^k : \mathrm{Alt}^{r+2k}(T_xM \otimes_{\mathbb{R}} \mathbb{C}) \to \mathrm{Alt}^r(T_xM \otimes_{\mathbb{R}} \mathbb{C}).$$

As L_x^k is surjective, each K_x^k is injective and we get a quasi-isometry

$$K : L^2\Omega^{r+2k}(M) \to L^2\Omega^r(M)$$

which is the adjoint of L. As K is injective, L has dense image. As L is a quasi-isometry, L has closed image. Hence L^k is surjective. The proof for $\mathcal{H}^r(M)$ is analogous. □

COROLLARY 7.6. *Let M be a compact Kähler manifold with $n = \dim_{\mathbb{R}}(M)$, $m = \dim_{\mathbb{C}}(M)$. Put*

$$b_r(M) = \dim_{\mathbb{Q}}\bigl(H_r(M,\mathbb{Q})\bigr) = \dim_{\mathbb{C}}\bigl(\mathcal{H}^r(M)\bigr) \quad \text{and}$$
$$h_{p,q}(M) = \dim_{\mathbb{C}}\bigl(\mathcal{H}^{p,q}(M)\bigr).$$

Then:
(1) $b_r(M) = \sum_{p+q=r} h_{p,q}(M)$;
(2) $b_r(M) = b_{n-r}(M)$, $h_{p,q}(M) = h_{m-p,m-q}(M)$;
(3) $h_{p,q}(M) = h_{q,p}(M)$;
(4) $b_r(M)$ *is even for r odd*;
(5) $h_{1,0}(M) = \frac{1}{2}\cdot b_1(M)$ *depends only on* $\pi_1(M)$;
(6) $b_r(M) \leqslant b_{r+2}(M)$ *for* $r \leqslant m$.

The following definitions are taken from [108].

DEFINITION 7.7. A p-form η is *bounded* if

$$\|\eta\|_\infty = \sup\bigl\{\|\eta_x\| \mid x \in M\bigr\} < \infty.$$

A p-form ω is *d(bounded)* if $\omega = d\eta$ for a bounded $(p-1)$-form η. We call ω *$\widetilde{d}$(bounded)* if its lift $\widetilde{\omega}$ to $\widetilde{M}$ is d(bounded).

DEFINITION 7.8. A *Kähler hyperbolic manifold M* is a compact Kähler manifold M whose Kähler form is $\widetilde{d}$(bounded).

THEOREM 7.9 (Gromov [108, Main Theorem 2.5, p. 283]). *Let M be a Kähler hyperbolic manifold with $n = 2m = \dim_{\mathbb{R}}(M)$ and universal covering $\widetilde{M}$. Then*
(1) *We get outside the middle dimension*:

$$\mathcal{H}^{p,q}_{(2)}(\widetilde{M}) = 0 \quad \text{for } p+q \neq m;$$
$$b^{(2)}_r(\widetilde{M}) = 0 \quad \text{for } r \neq m;$$

(2) *We get for the middle dimension*:

$$\mathcal{H}^{p,q}_{(2)}(\widetilde{M}) \neq 0 \quad \text{for } p+q = m;$$
$$b^{(2)}_m(\widetilde{M}) \neq 0;$$

(3) $(-1)^m \cdot \chi(M) > 0$.

PROOF. We only give the proof of the easy part (1). Notice that (3) follows from (1) and (2) by the Euler–Poincaré formula 1.7.2. Because of Poincaré duality and the L^2-Hodge–deRham decomposition 1.7, it suffices to prove

$$\mathcal{H}^r_{(2)}(\widetilde{M}) = 0 \quad \text{for } r < m.$$

Choose k with $2r + 2k = n$. Let $\widetilde{\omega}$ be the lift of the Kähler form ω on M to the universal covering $\widetilde{M}$. Then by the L^2-Lefschetz Theorem 7.5

$$L^k : \mathcal{H}^r_{(2)}(\widetilde{M}) \to \mathcal{H}^{r+2k}_{(2)}(\widetilde{M}), \qquad \phi \mapsto \phi \wedge \widetilde{\omega}^k,$$

is bijective. Put $\mu = \phi \wedge \widetilde{\omega}^{k-1} \wedge \eta$ for $\widetilde{\omega} = d\eta$ with $\|\eta\|_\infty < \infty$. As $\|\omega^{k-1}\|_\infty < \infty$, μ is an L^2-form. Obviously $L^k(\phi) = d(\mu)$. By the L^2-Hodge–deRham decomposition Theorem 7.4, $L^k(\phi) = 0$ and hence $\phi = 0$. □

Theorem 7.9 has been generalized by Jost and Zuo [130]. Further information about L^2-cohomology and Kähler manifolds can be found in [1].

Next we make some comments about the notion of Kähler hyperbolicity and applications of Theorem 7.9. Here is a list of examples of Kähler hyperbolic manifolds (see [108, Section 0]):

- M is a compact Kähler manifold and homotopy equivalent to a compact Riemannian manifold of negative sectional curvature;
- M is a compact Kähler manifold, $\pi_1(M)$ is hyperbolic in Gromov's sense [107] and $\pi_2(M) = 0$;
- $\widetilde{M}$ is a symmetric Hermitian space of non-compact type with no Euclidean factor;
- M is a submanifold of a Kähler hyperbolic manifold;
- M is a product of two Kähler hyperbolic manifolds.

Next we give two compact Kähler manifolds which are not Kähler hyperbolic. Equip $\mathbb{C}P^n$ with the Fubini-study metric which is, up to a positive constant, uniquely determined by the property that it is $U(n+1)$-invariant. Then the fundamental 2-form is the first Chern class $c_1(L)$. This is closed but not exact. Hence $\mathbb{C}P^n$ is Kähler but not Kähler hyperbolic. The torus T^{2n} with a complex structure is Kähler but cannot be Kähler hyperbolic because $\chi(T^{2n}) = 0$.

THEOREM 7.10. *Let M be a compact Kähler manifold. Then the following assertions are equivalent and they are true if M is Kähler hyperbolic*:

(1) *M is Moishezon, i.e., the transcendental degree of the field of meromorphic functions on M is* $\dim_{\mathbb{C}}(M)$;
(2) *M is Hodge, i.e., the Kähler form represents an integral cohomology class, i.e., it represents an element in the image of $H^2(M, \mathbb{Z}) \to H^2(M, \mathbb{C})$*;
(3) *M can be holomorphically embedded in $\mathbb{C}P^n$*;
(4) *M is a projective algebraic variety.*

PROOF. The equivalence of these statements is due to Kodaira [134], [249, Chapter VI] and Moishezon [177]. The first assertion is a consequence of Theorem 7.9 and proven in [108, Section 3] using a version of the L^2-index theorem. □

8. Novikov–Shubin invariants

In this section we introduce and study spectral density functions and Novikov–Shubin invariants. They were introduced analytically by Novikov and Shubin in [196].

DEFINITION 8.1. Let $f : U \to V$ be a morphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Let $\{E_\lambda^{f^*f} \mid \lambda \in \mathbb{R}\}$ denote the (right-continuous) family of spectral projections of the positive operator f^*f. Define the *spectral density function of* f by

$$F(f,\lambda) : \mathbb{R} \to [0,\infty), \qquad \lambda \mapsto \dim_{\mathcal{N}(\Gamma)}\bigl(\operatorname{im}\bigl(E_{\lambda^2}^{f^*f}\bigr)\bigr).$$

The spectral density function is monotone and right-continuous. It takes values in $[0, \|f\|]$. Here, and in the sequel, $|x|$ denotes the norm of an element x of a Hilbert $\mathcal{N}(\Gamma)$-module and $\|f\|$ the operator norm of a morphism. Since f and f^*f have the same kernel, $\dim_{\mathcal{N}(\Gamma)}(\ker(f)) = F(f,0)$. Given two morphisms f and g, we call their spectral density functions *dilatationally equivalent* if there are constants $C > 0$ and $\varepsilon > 0$ such that

$$F\bigl(f, C^{-1}\cdot\lambda\bigr) \leqslant F(g,\lambda) \leqslant F(f, C\cdot\lambda) \quad \text{for } \lambda \leqslant \varepsilon$$

holds.

EXAMPLE 8.2. Suppose that Γ is finite. Then a morphism $f : U \to V$ of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules is just a linear Γ-equivariant map of (over $\mathbb{C}$ finite-dimensional) unitary Γ-representations. Let $0 \leqslant \lambda_0 < \cdots < \lambda_r$ be the eigenvalues of f^*f and μ_i be the multiplicity of λ_i, i.e., the dimension of the eigenspace of λ_i. Then the spectral density function is a right continuous step function which is zero for $\lambda < 0$ and has a step of height $\mu_i/|\Gamma|$ at each $\sqrt{\lambda_i}$. Given two such maps f and g, their spectral density functions are dilatationally equivalent if and only if the kernels of f and g have the same complex dimension.

These notions become much more interesting in the case when Γ is infinite, because then the spectrum of f^*f is in general not discrete anymore.

LEMMA 8.3.

(1) *Let $f : U \to V$ be a morphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Let $\mathcal{L}(f,\lambda)$ denote the set of all Hilbert $\mathcal{N}(\Gamma)$-submodules L of U with the property that $|f(x)| \leqslant \lambda\cdot|x|$ holds for $x \in L$. Then*

$$F(f,\lambda) = \sup\bigl\{\dim_{\mathcal{N}(\Gamma)}(L) \mid L \in \mathcal{L}(f,\lambda)\bigr\};$$

(2) *Let $f : U \to V$ and $g : V \to W$ be morphisms of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Then*

$$F(f,\lambda) \leqslant F\bigl(gf, \|g\|\cdot\lambda\bigr).$$

If additionally f has dense image, then

$$F(g,\lambda) \leqslant F\big(gf, \|f\| \cdot \lambda\big);$$

(3) *Let $u: U \to V$, $f: V \to W$ and $i: W \to X$ be morphisms of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Suppose that u is an isomorphism and i is injective with closed image. Then the spectral density functions of f and $i \circ f \circ u$ are dilatationally equivalent.*

(4) *Let $f_i: U_i \to V_i$ be morphisms of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules for $i = 0, 1$. Then*

$$F(f_0 \oplus f_1, \lambda) = F(f_0, \lambda) + F(f_1, \lambda).$$

PROOF. (1) If $E_{\lambda^2}^{f^*f}(x) = x$,

$$\left|f(x)\right|^2 = \left|\int_0^\infty \mu \, \mathrm{d}\langle E_\mu^{f^*f}(x), x\rangle\right| \leqslant \lambda^2 \cdot \left|\int_0^\infty 1 \, \mathrm{d}\langle E_\mu^{f^*f}(x), x\rangle\right| \leqslant \lambda^2 \cdot |x|^2.$$

Hence the image of $E_{\lambda^2}^{f^*f}$ belongs to $\mathcal{L}(f, \lambda)$. This shows

$$F(f,\lambda) \leqslant \sup\big\{\dim_{\mathcal{N}(\Gamma)}(L) \mid L \in \mathcal{L}(f,\lambda)\big\}.$$

It remains to prove that $\dim_{\mathcal{N}(\Gamma)}(L) \leqslant \dim_{\mathcal{N}(\Gamma)}(\operatorname{im}(E_{\lambda^2}^{f^*f}))$ holds for all $L \in \mathcal{L}(f, \lambda)$. If $\lambda \geqslant 0$ and $x \in U$ satisfies $E_{\lambda^2}^{f^*f}(x) = 0$ and $x \neq 0$, then $|f(x)| > \lambda \cdot |x|$. Hence $E_{\lambda^2}^{f^*f}$ induces a weak isomorphism from L to $\operatorname{clos}(E_{\lambda^2}^{f^*f}(L))$ and the claim follows from Lemma 1.4.

(2) Consider $L \in \mathcal{L}(f, \lambda)$. For all $x \in L$ we get $|gf(x)| \leqslant \|g\| \cdot |f(x)| \leqslant \|g\| \cdot \lambda \cdot |x|$. This implies that $L \in \mathcal{L}(gf, \|g\| \cdot \lambda)$, and the first equation follows.

Now suppose that f has dense image. Consider $L \in \mathcal{L}(g, \lambda)$. For all $x \in f^{-1}(L)$, we have

$$\left|gf(x)\right| \leqslant \lambda \cdot \left|f(x)\right| \leqslant \lambda \cdot \|f\| \cdot |x|.$$

This implies $f^{-1}(L) \in \mathcal{L}(gf, \|f\| \cdot \lambda)$. It remains to show

$$\dim_{\mathcal{N}(\Gamma)}(L) \leqslant \dim_{\mathcal{N}(\Gamma)}\big(f^{-1}(L)\big).$$

Let $p: U \to U/\ker f$ be the projection and let $\bar{f}: U/\ker(f) \to V$ be the map induced by f. Since p is surjective, we get from Lemma 1.4

$$\dim_{\mathcal{N}(\Gamma)}\big(f^{-1}(L)\big) \geqslant \dim_{\mathcal{N}(\Gamma)}\big(p(f^{-1}(L))\big) = \dim_{\mathcal{N}(\Gamma)}\big(\bar{f}^{-1}(L)\big).$$

Next we show that the weak isomorphism $\bar{f}$ induces a weak isomorphism from $\bar{f}^{-1}(L)$ to L. Notice that then $\dim_{\mathcal{N}(\Gamma)}(\bar{f}^{-1}(L)) = \dim_{\mathcal{N}(\Gamma)}(L)$ because of Lemma 1.4, and the second equation will follow.

Because of the Polar Decomposition Theorem applied to $\bar{f}$, it suffices to prove for a positive weak isomorphism $h : V \to V$ and a Hilbert $\mathcal{N}(\Gamma)$-submodule $L \subset V$ that $h(h^{-1}(L))$ is dense in L. Now L has an orthogonal decomposition of the form $L = \operatorname{clos}(h(h^{-1}(L))) \oplus M$, where M is a $\mathcal{N}(\Gamma)$-submodule of L. If we can show $\dim_{\mathcal{N}(\Gamma)} M = 0$, then Lemma 1.4.1 will imply $M = 0$ and we will be done. As $h(h^{-1}(M)) \subset M$ and $h(h^{-1}(M)) \subset h(h^{-1}(L))$, it follows that $h(h^{-1}(M)) = 0$. Therefore $M \cap \operatorname{im}(h) = 0$. For $\lambda > 0$, consider the map $\pi_\lambda : M \to E_\lambda^h(V)$ given by $\pi_\lambda(m) = E_\lambda^h(m)$. If $m \in \ker(\pi_\lambda)$, then the spectral theorem shows that $m \in \operatorname{im}(h)$. Therefore $\ker(\pi_\lambda) = 0$, and Lemma 1.4 implies

$$\dim_{\mathcal{N}(\Gamma)} M \leqslant \dim_{\mathcal{N}(\Gamma)}\bigl(E_\lambda^h(V)\bigr).$$

As h is injective, continuity of the dimension 1.4.3 and the right-continuity of the spectral family implies

$$\lim_{\lambda \to 0^+} \dim_{\mathcal{N}(\Gamma)}\bigl(E_\lambda^h(V)\bigr) = \dim_{\mathcal{N}(\Gamma)}\bigl(E_0^h(V)\bigr) = \dim_{\mathcal{N}(\Gamma)}\bigl(\ker(h)\bigr) = 0.$$

Hence $\dim_{\mathcal{N}(\Gamma)} M = 0$ and assertion (2) is proven.

(3) By the Open Mapping Theorem there is a constant $D > 0$ such that

$$D^{-1} \cdot |x| \leqslant |i(x)| \leqslant D \cdot |x|$$

holds for all $x \in W$. Hence $F(i \circ f \circ u)$ and $F(f \circ u)$ are dilatationally equivalent by the first assertion. By the second assertion, $F(f \circ u)$ and $F(f)$ are dilatationally equivalent.

(4) This follows from additivity of the dimension 1.4.4. □

DEFINITION 8.4. Let $f : U \to V$ be a morphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Define its *Novikov–Shubin invariant* by

$$\alpha(f) = \liminf_{\lambda \to 0^+} \frac{\ln(F(f,\lambda) - F(f,0))}{\ln(\lambda)} \in [0, \infty],$$

provided that $F(f,\lambda) > F(f,0)$ holds for all $\lambda > 0$. Otherwise, we put

$$\alpha(f) = \infty^+.$$

Here ∞^+ is a new formal symbol which should not be confused with ∞.

Let $\overline{X} \to X$ be a regular covering over the CW-complex X of finite type with Γ as group of deck transformations. Define its *pth Novikov–Shubin invariant*

$$\alpha_p(\overline{X}) = \alpha(c_p^{(2)}) \in [0, \infty] \amalg \{\infty^+\},$$

where $c_p^{(2)}$ is the pth differential in the L^2-chain complex of $\overline{X}$ introduced in Definition 1.5.

EXAMPLE 8.5. We consider the example of the universal covering of S^1 with $\mathbb{Z}$ as group of deck transformations using the notation and results of Example 1.8 and Example 3.1.

The spectral family of the first differential $c_1^{(2)}$ of the L^2-chain complex has as projection for λ the operator given by multiplication with the characteristic function of the set $\{z \in S^1 \mid |z-1| \leqslant \lambda\}$. Hence for small $\lambda > 0$ we get for the spectral density function

$$\begin{aligned} F\big(c_1^{(2)}, \lambda\big) &= \operatorname{vol}\big\{z \in S^1 \mid |z-1| \leqslant \lambda\big\} \\ &= \operatorname{vol}\big\{\cos(\phi) + i \cdot \sin(\phi) \mid \lambda \geqslant \big|2 - 2\cos(\phi)\big|\big\}. \end{aligned}$$

Because of

$$\lim_{\phi \to 0} \frac{2 - 2\cos(\phi)}{\phi^2} = 1$$

$F(c_1^{(2)}, \Lambda)$ and λ are dilatationally equivalent, and hence we get

$$\alpha_1\big(\widetilde{S^1}\big) = 1.$$

Notice that $\alpha(f) = \infty^+$ precisely if and only if $f^* f$ has a gap in the spectrum above 0. Moreover, a morphism $f : U \to V$ of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules is an isomorphism if and only if $\dim_{\mathcal{N}(\Gamma)}(\ker(f)) = 0$, $\dim_{\mathcal{N}(\Gamma)}(U) = \dim_{\mathcal{N}(\Gamma)}(V)$ and $\alpha(f) = \infty^+$. If there is no gap in the spectrum above 0, then $\alpha(f)$ is the supremum over all non-negative numbers β for which there is an $\varepsilon > 0$ such that

$$F(f, \lambda) - F(f, 0) \leqslant t\lambda^\beta \quad \text{for } 0 \leqslant \lambda \leqslant \varepsilon.$$

The invariant $\alpha(f)$ measures how fast $F(f, \lambda)$ approaches $F(f, 0)$ for $\lambda \to 0+$.

REMARK 8.6. Notice that the Novikov–Shubin invariant $\alpha_p(\overline{X})$ is ∞^+ if and only if the image of the pth differential $c_p^{(2)}$ of the L^2-chain complex of $\overline{X}$ is closed. Hence $\alpha_p(\overline{X})$ measures the difference between the L^2-homology of $\overline{X}$, which is $\ker(c_p^{(2)})/\operatorname{clos}(\operatorname{im}(c_{p+1}^{(2)}))$, and the Γ-equivariant homology of $\overline{X}$ with coefficients in the $\mathbb{Z}\Gamma$-module $l^2(\Gamma)$, which is $\ker(c_p^{(2)})/\operatorname{im}(c_{p+1}^{(2)})$. We will continue this discussion in Section 10.

Before we give the main properties of the Novikov–Shubin invariants, we recall some notions and facts from group theory. A finitely generated group Γ is *nilpotent* if Γ possesses a finite lower central series

$$\Gamma = \Gamma_1 \supset \Gamma_2 \supset \cdots \supset \Gamma_s = \{1\}, \quad \Gamma_{k+1} = [\Gamma, \Gamma_k].$$

If $\overline{\Gamma}$ contains a nilpotent subgroup Γ of finite index, then $\overline{\Gamma}$ is said to be *virtually nilpotent.* Let d_i be the rank of the quotient Γ_i/Γ_{i+1} and let d be the integer $\sum_{i \geqslant 1} i d_i$. Then $\overline{\Gamma}$ has polynomial growth of degree d [11]. Note that a group has polynomial growth if and only if it is virtually nilpotent [105].

THEOREM 8.7.

(1) *Homotopy invariance*
Let $\overline{X}$ and $\overline{Y}$ be regular coverings of CW-complexes X and Y of finite type with the same group Γ of deck transformations. Let $f:\overline{X}\to\overline{Y}$ be a Γ-equivariant map. If f is a homotopy equivalence, then the spectral density functions of $c_p^{(2)}$ for $\overline{X}$ and $\overline{Y}$ are dilatationally equivalent for all p, and in particular

$$\alpha_p(\overline{X})=\alpha_p(\overline{Y}) \quad \textit{for } 0\leqslant p.$$

If f is d-connected, i.e., f induces an isomorphism on π_n for $n<d$ and an epimorphism on π_d, then the spectral density functions of $c_p^{(2)}$ for $\overline{X}$ and $\overline{Y}$ are dilatationally equivalent for $p\leqslant d$, and in particular

$$\alpha_p(\overline{X})=\alpha_p(\overline{Y}) \quad \textit{for } p\leqslant d;$$

(2) *Equality of analytic and combinatorial version*
Let $\overline{M}$ be a covering of the oriented closed Riemannian manifold M with deck transformation group Γ. We can define the spectral density function (see Example 1.9) and Novikov–Shubin invariants analytically in terms of the Laplace operator acting on differential forms on $\overline{M}$. Then the analytic and the combinatorial spectral density function are dilatationally equivalent. In particular, the analytically defined Novikov–Shubin invariants and the combinatorial Novikov–Shubin invariants agree.

(3) *Poincaré duality*
Let $\overline{M}$ be a regular covering of the closed manifold M of dimension n. Then

$$\alpha_p(\overline{M})=\alpha_{n+1-p}(\overline{M});$$

(4) *Dependency on the fundamental group*
The Novikov–Shubin invariants of the universal covering of a connected CW-complex of finite type $\alpha_p(\widetilde{X})$ for $p\leqslant 2$ depend only on $\pi_1(X)$. If M is a closed n-dimensional manifold with $n\leqslant 4$, then $\alpha_p(\widetilde{M})$ depends only on $\pi_1(M)$ for all p;

(5) *Invariance under finite coverings*
Let X be a CW-complex of finite type and $p:\overline{X}\to X$ be a regular covering with group of deck transformations Γ. Let $\Gamma_0\subset\Gamma$ be a subgroup of Γ of finite index n. We obtain a regular covering denoted by $\overline{\overline{X}}$ by $\overline{X}\to\overline{X}/\Gamma_0$. Notice that the coverings $\overline{\overline{X}}$ and $\overline{X}$ have the same total spaces but different groups of deck transformations. Then

$$\alpha_p(\overline{\overline{X}})=\alpha_p(\overline{X}) \quad \textit{for } p\geqslant 0;$$

(6) *First Novikov–Shubin invariant*
Let X be a connected CW-complex of finite type with fundamental group π and universal covering $\widetilde{X}$. Then

(a) $\alpha_1(\widetilde{X})$ *is finite if and only if* π *is infinite and virtually nilpotent. In this case* $\alpha_1(\widetilde{X})$ *is the growth rate of* π;

(b) $\alpha_1(\widetilde{X})$ *is* ∞^+ *if and only if* π *is finite or nonamenable;*

(c) $\alpha_1(\widetilde{X})$ *is* ∞ *if and only if* π *is amenable and not virtually nilpotent;*

(7) *S^1-actions and Novikov–Shubin invariants*
Let M be a connected closed manifold with S^1-action. Suppose that for one orbit S^1/H (and hence all orbits) the inclusion into M induces a map on π_1 with infinite image. (In particular, the S^1-action has no fixed points.) Then

$$\alpha_p(\widetilde{M}) \geqslant 1 \quad \textit{for all } p;$$

(8) *Positivity of the Novikov–Shubin invariants for* 3*-manifolds*
Let M be a 3*-manifold which has finite fundamental group or satisfies the assumptions of Theorem* 3.3. *Then we get for the Novikov–Shubin invariants of the universal covering*

$$\alpha_p(\widetilde{M}) > 0 \quad \textit{for all } p;$$

(9) *Novikov–Shubin invariants for $\mathbb{Z}$ as group of deck transformations*
Let $\overline{X} \to X$ be a regular covering of the CW-complex X of finite type with $\mathbb{Z}$ as group of deck transformations. Since $\mathbb{C}[\mathbb{Z}]$ is a principal ideal domain, one can write the $\mathbb{C}[\mathbb{Z}]$-module

$$H_{p-1}(\overline{X}, \mathbb{C}) = \mathbb{C}[\mathbb{Z}]^n \oplus \bigoplus_{i=1}^{k} \mathbb{C}[\mathbb{Z}]/\big((z-a_i)^{r_i}\big)$$

for integers n, k and r_i with $n, k \geqslant 0$ and $r_i \geqslant 1$ and $a_i \in \mathbb{C}$ with $a_i \neq 0$ and $a_i \neq a_j$ for $i \neq j$, where $z \in \mathbb{Z}$ is the generator. Then

$$\alpha_p(\overline{X}) = \min\left\{ \frac{1}{r_i} \,\middle|\, i = 1, 2 \ldots, r,\ a_i \in S^1 \right\}$$

if $k \geqslant 1$ and there is at least one a_i with $a_i \in S^1$, and

$$\alpha_p(\overline{X}) = \infty^+$$

otherwise;

(10) *Hyperbolic manifolds*
If M is a hyperbolic closed manifold of dimension n, then

$$\alpha_p(\overline{M}) = 1 \qquad \textit{if } n \textit{ is odd and } p = \frac{n \pm 1}{2}, \textit{ and}$$
$$\alpha_p(\overline{M}) = \infty^+ \quad \textit{otherwise;}$$

(11) *Kähler hyperbolic manifolds*
If M is Kähler hyperbolic in the sense of Definition 7.8, *then*

$$\alpha_p(\widetilde{M}) = \infty^+ \quad \textit{for } p \geqslant 0.$$

PROOF. (1) As in the proof of Theorem 1.7.1, one justifies the assumption that f is a cellular Γ-homotopy equivalence. Hence it suffices to show, for a chain map $f : C \to D$ of chain complexes of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules, that the spectral density functions of the pth differentials c_p and d_p are dilatationally equivalent for all p.

We begin with the case $D = 0$. This means that C is contractible and we must prove $\alpha_p(C) = \infty^+$. Let γ_* be a chain contraction for C, i.e., a collection of morphisms $\gamma_p : C_p \to C_{p+1}$ satisfying $c_{p+1} \circ \gamma_p + \gamma_{p-1} \circ c_p = \mathrm{id}$. Using c_p and γ_{p-1}, we can construct morphisms

$$\overline{c_p} : C_p / \operatorname{clos}(\operatorname{im}(c_{p+1})) \to C_{p-1} \quad \text{and} \quad \overline{\gamma_{p-1}} : C_{p-1} \to C_p / \operatorname{clos}(\operatorname{im}(c_{p+1}))$$

such that $\overline{\gamma_{p-1}} \circ \overline{c_p} = \mathrm{id}$. Hence $\overline{c_p}$ induces an invertible operator onto its image. Lemma 8.3.3 implies $\alpha(c_p) = \alpha(\overline{c_p}) = \infty^+$.

The case where f is an isomorphism of chain complexes follows from Lemma 8.3.3. Now we can treat the general case.

There are exact sequences of chain complexes $0 \to C \to \mathrm{cyl}(f) \to \mathrm{cone}(f) \to 0$ and $0 \to D \to \mathrm{cyl}(f) \to \mathrm{cone}(C) \to 0$, where cone denotes the mapping cone. The chain complexes $\mathrm{cone}(f)$ and $\mathrm{cone}(C)$ are contractible as chain complexes of Hilbert $\mathcal{N}(\Gamma)$-modules since f is a homotopy equivalence by assumption. We obtain chain isomorphisms $C \oplus \mathrm{cone}(f) \to \mathrm{cyl}(f)$ and $D \oplus \mathrm{cone}(C) \to \mathrm{cyl}(f)$ by the following general construction for an exact sequence $0 \to C \xrightarrow{j} D \xrightarrow{q} E \to 0$ with contractible E: Choose a chain contraction ε for E, and for each p a morphism $t_p : E_p \to D_p$ such that $q_p \circ t_p = \mathrm{id}$. Put

$$s_p = d_{p+1} \circ t_{p+1} \circ \varepsilon_p + t_p \circ \varepsilon_{p-1} \circ e_p.$$

This defines a chain map $s : E \to D$ such that $q \circ s = \mathrm{id}$. Define a chain map $u : D \to C$ by mapping $x \in D_p$ to $y = u_p(x)$ which is the unique element $y \in C_p$ such that $x = s_p q_p(x) + j_p(y)$. Then $j + s$ is a chain isomorphism $C \oplus E \to D$ with inverse $u \oplus q$. Since $C \oplus \mathrm{cone}(f)$ and $D \oplus \mathrm{cone}(C)$ are isomorphic and $\mathrm{cone}(f)$ and $\mathrm{cone}(C)$ are contractible, the claim follows from the special cases which we have already proven and Lemma 8.3.4.

The more general case of a d-connected map is proven in [148, Lemma 3.3, p. 33]. The homotopy invariance of the analytically defined Novikov–Shubin invariants is proven by Gromov and Shubin [110].

(2) This is proven by Efremov [84,85].
(3) This follows from homotopy invariance as in the proof of Theorem 1.7.3.
(4) This follows from (1) applied to the classifying map $\widetilde{X} \to E\Gamma$ and Poincaré duality.
(5) This is similar to the proof of Theorem 8.3.7.
(6) This is proven in [148, Lemma 3.5, p. 34] using [34] and [241].
(7) This is proven in [148, Theorem 3.1].

(8) This is proven in [148, Theorem 0.1].
(9) This is proven in [157, Example 4.3].
(10) This is proven in [142, Proposition 46, p. 499] and [75].
(11) This is proven in [108, Theorem 1.4.A, p. 274]. □

REMARK 8.8. We have already mentioned in Remark 1.9 that the L^2-Betti numbers are invariants of the asymptotic large time behaviour of the heat kernel on the regular covering $\overline{M} \to M$ of a closed Riemannian manifold M (see (1.12)). The Novikov–Shubin invariants measure the speed of convergence of the limit as $t \to \infty$ in the analytic definition of the L^2-Betti numbers. Namely, one gets

$$\min\{\alpha_p(\overline{M}), \alpha_{p-1}(\overline{M})\} = \sup\left\{\beta_p \in [0,\infty) \;\middle|\; \lim_{t\to\infty} \frac{\int_{\mathcal{F}} \operatorname{tr}_{\mathbb{C}}(e^{-t\Delta_p}(\overline{x},\overline{x}))\,d\overline{x} - b_p^{(2)}(\overline{M})}{t^{\beta_p/2}} = 0\right\}$$

in $[0,\infty]$ where we do not distinguish between ∞ and ∞^+ here. The supremum of the right is the same as the supremum over all numbers $\beta_p \geqslant 0$ for which there is a $K > 0$ such that

$$\int_{\mathcal{F}} \operatorname{tr}_{\mathbb{C}}\big(e^{-t\Delta_p}(\overline{x},\overline{x})\big)\,d\overline{x} - b_p^{(2)}\big(\overline{M}\big) \leqslant t^{\beta_p/2} \quad \text{for } K \leqslant t.$$

See the discussion in [110, appendix] about decay exponents and Laplace transforms.

We mention that more explicit calculations of the Novikov–Shubin invariants for 3-manifolds can be found in [148]. The connection for abelian fundamental groups with Massey products and information about locally symmetric spaces can be found in [142, Section VI and VII]. Finally we mention the following conjecture [148, Conjecture 7.1, p. 56].

CONJECTURE 8.9 (*Rationality and Positivity of Novikov–Shubin invariants*). The Novikov–Shubin invariants of the universal covering $\widetilde{M}$ of a closed Riemannian manifold M are all positive and rational.

Conjecture 8.9 is true if M is hyperbolic or Kähler hyperbolic, or if the fundamental group of M is abelian or free. The positivity of the Novikov–Shubin invariants plays a role in the definition of L^2-torsion of Section 9 and is proven in [148, Theorem 0.1, p. 16] for all 3-manifolds satisfying the assumption of Theorem 3.3 that none of its prime factors is exceptional. A similar discussion as for Conjecture 2.1, Lemma 2.2 and Remark 2.3 can be made for the Conjecture 8.9 above [148, Section 7]. Novikov–Shubin invariants for arbitrary spaces with Γ-action (without any finiteness assumptions) are constructed in [162]. Further references on Novikov–Shubin invariants are [89,90,111,157].

9. L^2-torsion

In this section we introduce and study L^2-torsion. This is the L^2-version of the classical combinatorially defined Reidemeister torsion and the analytically defined Ray–Singer torsion. We will restrict ourselves in this section to the case of the universal covering $\widetilde{X} \to X$ of a connected finite CW-complex, for simplicity and due the fact that this is the most important case for applications. The general case of a regular covering is not much harder.

We begin with recalling the notions of combinatorial Reidemeister torsion and analytic Ray–Singer torsion in order to motivate the L^2-versions we will introduce later. A reader who is familiar with these concepts may skip this part.

DEFINITION 9.1. Let $\widetilde{X} \to X$ be the universal covering of a connected finite CW-complex with group of deck transformations $\Gamma = \pi_1(X)$. Let V be a unitary (finite-dimensional) Γ-representation. Let $C_*(\widetilde{X})$ be the cellular $\mathbb{Z}\Gamma$-chain complex. We obtain a $\mathbb{C}$-chain complex $V \otimes_{\mathbb{Z}\Gamma} C_*(\widetilde{X})$. Its chain modules inherit a Hilbert space structure from the cellular $\mathbb{Z}\Gamma$-basis and the inner product on V. Suppose that $H_p(V \otimes_{\mathbb{Z}\Gamma} C_*(\widetilde{X}))$ is trivial for all $p \geqslant 0$. Define the combinatorial Laplace operator of $\widetilde{X}$ with coefficients in V by

$$\Delta_p = c_{p+1} \circ c_p^* + c_{p-1}^* \circ c_p : V \otimes_{\mathbb{Z}\Gamma} C_p(\widetilde{X}) \to V \otimes_{\mathbb{Z}\Gamma} C_p(\widetilde{X}).$$

Then the *Reidemeister torsion* of $\widetilde{X}$ with coefficients in V is defined as

$$\rho(\widetilde{X}; V) = -\sum_{p \geqslant 0} (-1)^p \cdot p \cdot \ln(\det_{\mathbb{C}}(\Delta_p)).$$

The original definition uses chain contractions of $V \otimes_{\mathbb{Z}\Gamma} C(\widetilde{X})$. The calculation in the proof of [163, Lemma 7.12, p. 257] shows that the definition above agrees with the logarithm of the classical one. We use the logarithm because then the combinatorial version will coincide with the analytic one.

Reidemeister [214] introduced this invariant to classify lens spaces up to PL-homeomorphism (and up to diffeomorphism) using the work of Franz [94] (see also [59, Chapter V], [175, Section 3]). This classification of lens spaces was generalized by deRham [215] who proved that two orthogonal G-representations V and W are isometrically $\mathbb{R}G$-isomorphic if and only if their unit spheres are G-diffeomorphic (see also [149, Proposition 3.2, p. 478], [153, p. 317], [220, Section 3]).

The result of deRham does not hold in the topological category. Namely, there are non-linearly isomorphic G-representations V and W whose unit spheres are G-homeomorphic, by the work of Cappell and Shaneson [46] (see also [114]). However, if G has odd order, G-homeomorphic implies G-diffeomorphic for unit spheres in G-representations as shown by Hsiang and Pardon [124], and by Madsen and Rothenberg [167].

Let M be a closed Riemannian manifold with $\Gamma = \pi_1(M)$. Then one defines $\rho(\widetilde{M}; V)$ to be $\rho(\widetilde{X}; V)$ for any smooth triangulation X of M, provided that $H_p(V \otimes_{\mathbb{Z}\Gamma} C_*(\widetilde{X}))$ vanishes for all $p \geqslant 0$. The last condition ensures that the definition is independent of the choice of triangulation and depends only on the simple homotopy type of M and, in particular, only on the homeomorphism type of M. If this condition is not satisfied, one modifies the definition by a term which is essentially given by the torsion of the deRham isomorphism and which relates this homology (which is canonically isomorphic to the cohomology) to the finite-dimensional Hilbert space given by harmonic forms on M with coefficients in V. Then $\rho(\widetilde{M}; V)$ is independent of the choice of triangulation but depends on the Riemannian metric.

Ray and Singer [211] defined the analytic counterpart of Reidemeister torsion using a regularization of the zeta-function as follows. Let M be a closed Riemannian manifold with $\Gamma = \pi_1(M)$. Let $\Delta_p : \Omega^p(M; V) \to \Omega^p(M; V)$ be the Laplace operator acting on smooth p-forms on M with coefficients in the unitary (finite-dimensional) Γ-representation V. This is an essentially selfadjoint operator with discrete spectrum as M is compact. The *zeta-function* is defined by

$$\zeta_p(s) = \sum_{\lambda>0} \lambda^{-s},$$

where λ runs through the positive eigenvalues of Δ_p listed with multiplicity. The zeta-function is holomorphic for $\Re(s) > \dim(M)/2$ and has a meromorphic extension to $\mathbb{C}$ with no poles in 0 [232]. So its derivative for $s = 0$ is defined. Now the *Ray–Singer torsion* of M is defined by [211, Definition 1.6, p. 149] (our definition is the old one multiplied by the factor 2)

$$\rho(M; V) = \sum_{p\geqslant 0} (-1)^p \cdot p \cdot \frac{\mathrm{d}}{\mathrm{d}s} \zeta_p(s)\Big|_{s=0}.$$

The basic idea is that $\frac{\mathrm{d}}{\mathrm{d}s}\zeta_p(s)|_{s=0}$ is a generalization of the ordinary determinant $\det_{\mathbb{C}}$. Namely, if $f : V \to V$ is a positive linear automorphism of the finite-dimensional complex vector space V and $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the eigenvalues of f listed with multiplicity, then we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}\zeta_p(s)\Big|_{s=0} &= \frac{\mathrm{d}}{\mathrm{d}s}\sum_{i=1}^{r} \lambda^{-s}\Big|_{s=0} = \sum_{i=1}^{r}\big(-\ln(\lambda)\cdot\lambda^{-s}\big)\Big|_{s=0} \\ &= -\ln\left(\prod_{i=1}^{r}\lambda_i\right) = -\ln\big(\det_{\mathbb{C}}(f)\big). \end{aligned}$$

Ray and Singer conjectured that the analytic and combinatorial versions agree. This conjecture was proven independently by Cheeger [50] and Müller [184]. Manifolds with boundary and manifolds with symmetries, sum (= glueing) formulas and fibration formulas are treated in [67,149,153,165,242–244]. Non-unitary coefficient systems are studied in [25,26,186]. Further references are [18–24,32,40,69,88,95,96,101,135,136,143,181,187, 208,212,245].

The definition of combinatorial L^2-torsion is based on the notion of the determinant which we treat next.

DEFINITION 9.2. Let $f: U \to V$ be a morphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Let $F(f, \lambda)$ be the spectral density function of Definition 8.1 which is a monotone non-decreasing right-continuous function. Let $\mathrm{d}F$ be the unique measure on the Borel σ-algebra on $\mathbb{R}$ which satisfies $\mathrm{d}F(]a, b]) = F(b) - F(a)$ for $a < b$. Then define the *(generalized) Kadison–Fuglede determinant*

$$\det_{\mathcal{N}(\Gamma)}(f) \in [0, \infty)$$

by the positive real number

$$\det_{\mathcal{N}(\Gamma)}(f) = \exp\left(\int_{0+}^{\infty} \ln(\lambda)\,\mathrm{d}F\right)$$

if the Lebesgue integral $\int_{0+}^{\infty} \ln(\lambda)\,\mathrm{d}F$ converges to a real number and by 0 otherwise.

EXAMPLE 9.3. To illustrate this definition, we look at the example where Γ is finite. We essentially get the ordinary determinant $\det_{\mathbb{C}}$. Namely, we have computed the spectral density function for finite Γ in Example 8.2. Let $\lambda_1, \lambda_2, \dots, \lambda_r$ be the non-zero eigenvalues of f^*f with multiplicity μ_i. Then one obtains, if $\overline{f^*f}$ is the automorphism of the orthogonal complement of the kernel of f^*f induced by f^*f,

$$\begin{aligned}\det_{\mathcal{N}(\Gamma)}(f) &= \exp\left(\sum_{i=1}^{r} \frac{\mu_i}{|\Gamma|} \cdot \ln\left(\sqrt{\lambda_i}\right)\right) = \prod_{i=1}^{r} \lambda_i^{\mu_i/(2\cdot|\Gamma|)} \\ &= \det_{\mathbb{C}}\left(\overline{f^*f}\right)^{1/(2\cdot|\Gamma|)}.\end{aligned}$$

If f is an isomorphism we get

$$\det_{\mathcal{N}(\Gamma)}(f) = \left|\det_{\mathbb{C}}(f)\right|^{1/|\Gamma|}.$$

If f is an isomorphism, then Definition 9.2 of $\det_{\mathcal{N}(\Gamma)}(f)$ reduces to the classical notion due to Fuglede and Kadison [97]. The proof of the next two lemmas can be found in [154, Lemma 4.1, p. 94 and Lemma 4.2, p. 97].

LEMMA 9.4. *Let $f: M \to N$ be a morphism of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules. Then*

(1) *We have for $0 < \varepsilon \leqslant a$:*

$$\begin{aligned}\int_{\varepsilon}^{a} \ln(\lambda)\,\mathrm{d}F &= -\int_{\varepsilon}^{a} \frac{1}{\lambda} \cdot \left(F(\lambda) - F(0)\right)\mathrm{d}\lambda \\ &\quad + \ln(a) \cdot \left(F(a) - F(0)\right) - \ln(\varepsilon) \cdot \left(F(\varepsilon) - F(0)\right); \\ \int_{0+}^{a} \ln(\lambda)\,\mathrm{d}F &= \lim_{\varepsilon \to 0+} \int_{\varepsilon}^{a} \ln(\lambda)\,\mathrm{d}F; \\ \int_{0+}^{a} \frac{1}{\lambda} \cdot \left(F(\lambda) - F(0)\right)\mathrm{d}\lambda &= \lim_{\varepsilon \to 0+} \int_{\varepsilon}^{a} \frac{1}{\lambda} \cdot \left(F(\lambda) - F(0)\right)\mathrm{d}\lambda;\end{aligned}$$

(2) *If the Novikov–Shubin invariant satisfies* $\alpha(f) > 0$ *and* $a \geqslant \|f\|$, *then the integrals*

$$\int_{0+}^{\infty} \ln(\lambda)\,\mathrm{d}F$$

and

$$\ln(a)\cdot\big(F(a)-F(0)\big) - \int_{0+}^{a} \frac{1}{\lambda}\cdot\big(F(\lambda)-F(0)\big)\,\mathrm{d}\lambda$$

do converge to the same real number and we have

$$\det_{\mathcal{N}(\Gamma)}(f) > 0.$$

LEMMA 9.5. *Let* M *and* N *be finitely generated Hilbert* $\mathcal{N}(\Gamma)$*-modules. Let* $s, t : M \to M$, $u : M \to N$ *and* $v : N \to N$ *be morphisms. Suppose that* s, t *and* v *have trivial kernel. Then we get*

(1) $\det_{\mathcal{N}(\Gamma)}(s) = \det_{\mathcal{N}(\Gamma)}(s^*) = \sqrt{\det_{\mathcal{N}(\Gamma)}(s^*s)} = \sqrt{\det_{\mathcal{N}(\Gamma)}(ss^*)}$;

(2) *If* $0 \leqslant s$ *(i.e., s is a positive operator), then*

$$\lim_{\varepsilon\to 0+} \det_{\mathcal{N}(\Gamma)}(s + \varepsilon\cdot \mathrm{id}) = \det_{\mathcal{N}(\Gamma)}(s);$$

(3) *If* $0 \leqslant s \leqslant t$, *then*

$$\det_{\mathcal{N}(\Gamma)}(s) \leqslant \det_{\mathcal{N}(\Gamma)}(t);$$

(4) $\det_{\mathcal{N}(\Gamma)}(st) = \det_{\mathcal{N}(\Gamma)}(s)\cdot\det_{\mathcal{N}(\Gamma)}(t)$;

(5) $\det_{\mathcal{N}(\Gamma)}\left(\begin{smallmatrix} s & u \\ 0 & v\end{smallmatrix}\right) = \det_{\mathcal{N}(\Gamma)}(s)\cdot\det_{\mathcal{N}(\Gamma)}(v)$.

The notation of determinant class below is taken from [43].

DEFINITION 9.6. Let $\widetilde{X} \to X$ be the universal covering of a finite CW-complex X with Γ as group of deck transformations. We call $\widetilde{X}$ of *determinant class* if $\det(\Delta_p) > 0$ holds for all $p \geqslant 0$ where Δ_p is the combinatorial Laplace operator introduced in (1.6). If $\widetilde{X}$ is of determinant class, we define its L^2*-torsion* as

$$\rho^{(2)}(\widetilde{X}) = -\sum_{p\geqslant 0} (-1)^p\cdot p\cdot \ln\big(\det_{\mathcal{N}(\Gamma)}(\Delta_p)\big) \in \mathbb{R}.$$

Schick [229] has introduced the class $\mathcal{G}$ of groups which has the following properties. All amenable groups belong to $\mathcal{G}$. Moreover, if $1 \to \Delta \to \Gamma \to \pi \to 1$ is an extension, $\Delta \in \mathcal{G}$ and π is amenable, then $\Gamma \in \mathcal{G}$. If Γ is the direct limit or the inverse limit of a directed system $\{\Gamma_i \mid i \in I\}$ of groups with $\Gamma_i \in \mathcal{G}$ for all $i \in I$, then $\Gamma \in \mathcal{G}$. The class $\mathcal{G}$ is closed under free products. It is residually closed and, in particular, contains all residually finite groups.

LEMMA 9.7.

(1) *Let X be a connected finite CW-complex. If $\alpha_p(\widetilde{X}) > 0$ for all $p \geqslant 0$ or if $\pi_1(X)$ belongs to $\mathcal{G}$, then $\widetilde{X}$ is of determinant class;*

(2) *Let $f : X \to Y$ be a homotopy equivalence of finite connected CW-complexes. If $\widetilde{X}$ or $\widetilde{Y}$ is of determinant class, then both $\widetilde{X}$ and $\widetilde{Y}$ are of determinant class.*

PROOF. (1) If $\alpha_p(\widetilde{X}) > 0$, this follows from Lemma 9.4.2. If $\Gamma \in \mathcal{G}$, the claim is proven in [229, Theorem 1.14] (see also [41, appendix], [56], [156, Theorem 3.4.2, p. 476]).

(2) We get from Theorem 8.7 that the spectral density functions of the pth differential, and hence of the combinatorial Laplace operator Δ_p on the L^2-chain complex of $\widetilde{X}$ and of $\widetilde{Y}$, are dilatationally equivalent. Lemma 9.4.1 implies that Δ_p for $\widetilde{X}$ is of determinant class if and only if the one for $\widetilde{Y}$ is. □

Next we describe the favorite situation for L^2-torsion.

DEFINITION 9.8. Let $\widetilde{X} \to X$ be the universal covering of a connected finite CW-complex X with $\Gamma = \pi_1(X)$. We call $\widetilde{X}$ of *acyclic determinant class* if $\widetilde{X}$ is of determinant class and $b_p^{(2)}(\widetilde{X}) = 0$ holds for all $p \geqslant 0$. We call $\widetilde{X}$ *admissible* if $\alpha_p(\widetilde{X}) > 0$ and $b_p^{(2)}(\widetilde{X}) = 0$ holds for all $p \geqslant 0$.

Of course admissible implies of acyclic determinant class because of Lemma 9.4.2. If X is not connected, we mean by the phrase that $\widetilde{X}$ is of acyclic determinant class (respectively admissible) that the universal covering of each component of X has this property, and we write $\rho^{(2)}(\widetilde{X})$ for the sum of the L^2-torsions of the universal coverings of the components of X. This remark is relevant for the sum formula appearing in the next theorem.

The L^2-torsion for universal coverings of finite CW-complexes of acyclic determinant class behaves like a multiplicative Euler characteristic, as the following result illuminates.

THEOREM 9.9.

(1) *Homotopy invariance*
Let $f : X \to Y$ be a homotopy equivalence of connected finite CW-complexes with $\Gamma = \pi_1(X) = \pi_1(Y)$. Suppose that $\widetilde{X}$ or $\widetilde{Y}$ is of acyclic determinant class (respectively admissible). Let

$$\Phi_\Gamma : \mathrm{Wh}(\Gamma) \to \mathbb{R}^{>0}$$

be the map from the Whitehead group of Γ (see [59, §11], [175, p. 373]) *to the multiplicative group of positive real numbers which assigns to the class of an invertible (n,n)-matrix A over $\mathbb{Z}\Gamma$ the determinant $\det_{\mathcal{N}(\Gamma)}(f)$ of the isomorphism $f : l^2(\Gamma)^n \to l^2(\Gamma)^n$ which is induced by right multiplication with A. Let $\tau(f) \in \mathrm{Wh}(\Gamma)$ be the Whitehead torsion* [59, Chapter IV], [175, p. 377]. *Then both $\widetilde{X}$ and $\widetilde{Y}$ are of acyclic determinant class (respectively admissible) and we get*

$$\rho^{(2)}(\widetilde{Y}) - \rho^{(2)}(\widetilde{X}) = \ln(\Phi_\Gamma(\tau(f)));$$

(2) *Fundamental groups belonging to $\mathcal{G}$*
Let $f : X \to Y$ be a homotopy equivalence of connected finite CW-complexes. Suppose that $\Gamma = \pi_1(X) = \pi_1(Y) \in \mathcal{G}$ and that $b_p^{(2)}(\widetilde{X})$ or $b_p^{(2)}(\widetilde{Y})$ is trivial for all $p \geqslant 0$. Then both $\widetilde{X}$ and $\widetilde{Y}$ are of acyclic determinant class and

$$\rho^{(2)}(\widetilde{X}) = \rho^{(2)}(\widetilde{Y});$$

(3) *Sum formula*
Consider the pushout of finite CW-complexes such that j_1 is an inclusion of CW-complexes and j_2 is cellular

$$\begin{array}{ccc} X_0 & \xrightarrow{j_1} & X_1 \\ {\scriptstyle j_2}\downarrow & & \downarrow{\scriptstyle i_1} \\ X_2 & \xrightarrow[i_2]{} & X \end{array}$$

Assume that $\widetilde{X_0}$, $\widetilde{X_1}$ and $\widetilde{X_2}$ are of acyclic determinant class (respectively admissible) and that for $i = 0, 1, 2$ the map $\pi_1(X_i) \to \pi_1(X)$ induced by the inclusion is injective for all base points in X_i. Then $\widetilde{X}$ is of acyclic determinant class (respectively admissible) and we get

$$\rho^{(2)}(\widetilde{X}) = \rho^{(2)}(\widetilde{X_1}) + \rho^{(2)}(\widetilde{X_2}) - \rho^{(2)}(\widetilde{X_0});$$

(4) *Fibration formula*
Let $F \to E \to B$ be a fibration of connected finite CW-complexes. Suppose that $\widetilde{F}$ is of acyclic determinant class (respectively admissible) and the inclusion induces an injection $\pi_1(F) \to \pi_1(E)$. Then $\widetilde{E}$ is of acyclic determinant class (respectively admissible) and we get

$$\rho^{(2)}(\widetilde{E}) = \chi(B) \cdot \rho^{(2)}(\widetilde{F})$$

where $\chi(B)$ is the Euler characteristic of B;

(5) *Product formula*
Let X and Y be connected finite CW-complexes. Suppose that $\widetilde{X}$ is of acyclic determinant class (respectively admissible). Then $\widetilde{X} \times \widetilde{Y}$ is of acyclic determinant class (respectively admissible) and we get

$$\rho^{(2)}(\widetilde{X} \times \widetilde{Y}) = \chi(Y) \cdot \rho^{(2)}(\widetilde{X});$$

(6) *Poincaré duality*
Let M be a closed manifold of even dimension. Suppose $\widetilde{M}$ is of acyclic determinant class. Then we get

$$\rho^{(2)}(\widetilde{M}) = 0;$$

(7) *Multiplicative property for finite coverings*
Let X be a connected finite CW-complex and $p: Y \to X$ be a finite d-sheeted covering. Suppose that $\widetilde{Y}$ or $\widetilde{X}$ is of acyclic determinant class (respectively admissible). Then both $\widetilde{Y}$ and $\widetilde{X}$ are of acyclic determinant class (respectively admissible) and we get

$$\rho^{(2)}(\widetilde{Y}) = d \cdot \rho^{(2)}(\widetilde{X});$$

(8) *S^1-actions*
Let M be a connected closed manifold with S^1-action. Suppose that for one orbit S^1/H (and hence all orbits) the inclusion into M induces a map on π_1 with infinite image. (In particular, the S^1-action has no fixed points.) Then $\widetilde{M}$ is admissible, and in particular, of acyclic determinant class and

$$\rho^{(2)}(\widetilde{M}) = 0.$$

PROOF. The proof is given for admissible CW-complexes in [156]. The case of determinant class follows analogously using the following result. Let $0 \to C \to D \to E \to 0$ be a short exact sequence of Hilbert $\mathcal{N}(\Gamma)$-chain complexes which are finite-dimensional and whose chain modules are finitely generated. Suppose for two of the chain complexes that they are L^2-acyclic and the associated Laplace operators are of determinant class in each dimension. Then the third chain complex has the same property. The statement about the acyclicity follows from the long weakly exact homology sequence of Cheeger and Gromov [52, Theorem 2.1, p. 10]. The strategy of proof for the determinant class is similar to the one in [148, Theorem 2.3, p. 27] using Lemma 9.5.5 instead of [148, Lemma 1.12, p. 25]. The case, where the fundamental group belongs to $\mathcal{G}$ follows from [229, Theorem 1.14]. □

We mention the following conjecture

CONJECTURE 9.10. The map $\Phi_\Gamma : \mathrm{Wh}(\Gamma) \to \mathbb{R}^{>0}$ defined in Theorem 9.9.1 is always trivial.

Suppose that M is a closed Riemannian manifold. Then one defines its L^2-torsion analogously to the classical definition of $\rho(\widetilde{M}; V)$. If the L^2-Betti numbers are not all trivial, one has to invoke a correction term analogously to the classical definition which involves the L^2-Hodge–deRham isomorphism. Details can be found, for instance, in [163, Section 5].

As for L^2-Betti numbers and for Novikov–Shubin invariants there are analytic versions of the L^2-torsion due to Lott [142] and Matthai [170].

DEFINITION 9.11. Let $\widetilde{M} \to M$ be the universal covering of a closed Riemannian manifold M with $\Gamma = \pi_1(M)$. Suppose that $\widetilde{M}$ is of determinant class. Define

$$\mathrm{tr}_{\mathcal{N}(\Gamma)}\left(e^{-t\Delta_p}\right) = \int_{\mathcal{F}} \mathrm{tr}_{\mathbb{C}}\left(e^{-t\Delta_p}(\widetilde{x}, \widetilde{x})\right) d\widetilde{x}$$

using the notation of Remark 1.9. Then we define the *analytic L^2-torsion of $\widetilde{M}$*

$$\rho^{(2)}(\widetilde{M}) = \sum_{p\geqslant 0}(-1)^p \cdot p \cdot \left(\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{\Gamma(s)}\int_0^{\varepsilon} t^{s-1} \times \left(\mathrm{tr}_{\mathcal{N}(\Gamma)}\left(\mathrm{e}^{-t\Delta_p}\right) - b_p^{(2)}(\widetilde{M})\right)\mathrm{d}t\bigg|_{s=0} + \int_{\varepsilon}^{\infty} t^{-1}\cdot\left(\mathrm{tr}_{\mathcal{N}(\Gamma)}\left(\mathrm{e}^{-t\Delta_p}\right) - b_p^{(2)}(\widetilde{M})\right)\mathrm{d}t\right).$$

The definition is independent of the choice of ε by the following calculation. The Γ-function

$$\Gamma(s) = \int_0^{\infty} t^{s-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$$

satisfies $\Gamma(s+1) = s\cdot\Gamma(s)$ and $\Gamma(1) = 1$. In the sequel we abbreviate

$$T(t) = \mathrm{tr}_{\mathcal{N}(\Gamma)}\left(\mathrm{e}^{-t\Delta_p}\right) - b_p^{(2)}(\widetilde{M}).$$

We compute for $0 < \varepsilon \leqslant \delta$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{\Gamma(s)}\int_{\varepsilon}^{\delta} t^{s-1}\cdot T(t)\,\mathrm{d}t\bigg|_{s=0} &= \frac{\mathrm{d}}{\mathrm{d}s}s\cdot\frac{1}{\Gamma(s+1)}\int_{\varepsilon}^{\delta} t^{s-1}\cdot T(t)\,\mathrm{d}t\bigg|_{s=0}\\ &= \frac{\mathrm{d}}{\mathrm{d}s}s\bigg|_{s=0}\cdot\frac{1}{\Gamma(s+1)}\int_{\varepsilon}^{\delta} t^{s-1}\cdot T(t)\,\mathrm{d}t\bigg|_{s=0}\\ &\quad + 0\cdot\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{\Gamma(s+1)}\int_{\varepsilon}^{\delta} t^{s-1}\cdot T(t)\,\mathrm{d}t\bigg|_{s=0}\\ &= \int_{\varepsilon}^{\delta} t^{-1}\cdot T(t)\,\mathrm{d}t.\end{aligned}$$

The following calculation shows the relation of the definition above with the classical Ray–Singer torsion. Namely, we get in the setting of the Ray–Singer torsion the following equation, where λ runs over the eigenvalues of the Laplace operator in dimension p listed with multiplicity:

$$\begin{aligned}\sum_{\lambda>0}\lambda^{-s} &= \sum_{\lambda>0}\frac{1}{\Gamma(s)}\cdot\lambda^{-s}\cdot\int_0^{\infty} t^{s-1}\,\mathrm{e}^{-t}\,\mathrm{d}t\\ &= \sum_{\lambda>0}\frac{1}{\Gamma(s)}\cdot\int_0^{\infty}\left(t\lambda^{-1}\right)^{s-1}\mathrm{e}^{-\lambda(t\lambda^{-1})}\lambda^{-1}\,\mathrm{d}t\\ &= \sum_{\lambda>0}\frac{1}{\Gamma(s)}\cdot\int_0^{\infty} t^{s-1}\,\mathrm{e}^{-\lambda t}\,\mathrm{d}t\end{aligned}$$

$$= \frac{1}{\Gamma(s)} \cdot \int_0^\infty t^{s-1} \cdot \sum_{\lambda>0} e^{-\lambda t}\, dt$$

$$= \frac{1}{\Gamma(s)} \cdot \int_0^\infty t^{s-1} \cdot \big(\mathrm{tr}_{\mathbb{C}}\big(e^{-t\Delta_p}\big) - \dim_{cc}\big(H_p(M; V)\big)\big)\, dt.$$

The integral from 0 to ε appearing in the Definition 9.11 exists by an argument analogous to the proof that the zeta-function is meromorphic without pole in 0 in the classical case. Given $p \geqslant 0$, the integral from ε to ∞ obviously converges if the Novikov–Shubin invariant $\alpha_p(\widetilde{M})$ is positive. It does converge if and only if for the analytic defined spectral density function F_p of the Laplace operator acting on p-forms on the universal covering (1.10) the integral $\int_{0+}^1 \ln(\lambda)\, dF$ does converge [43, Proposition 2.12]. Since the analytic and combinatorial spectral density functions are dilatationally equivalent by Theorem 8.7.2, $\int_{0+}^1 \ln(\lambda)\, dF$ converges if and only if $\widetilde{M}$ is of determinant class.

Next we mention the important result of Burghelea, Friedlander, Kappeler and McDonald which generalizes the Theorem of Cheeger and Müller to the L^2-case. The main technical tool is the generalization of the calculus of elliptic pseudo-differential operators, and of the Helffer–Sjöstrand analysis of the Witten deformation of the deRham complex for a closed Riemannian manifold with coefficients in a unitary finite-dimensional representation to a unitary representation on a finitely generated Hilbert $\mathcal{N}(\Gamma)$-module.

THEOREM 9.12 (Burghelea, Friedlander, Kappeler and McDonald [43]). *The analytic and the combinatorial L^2-torsion of the universal covering of a closed Riemannian manifold agree.*

Theorem 9.12 above allows us to combine the results we have already mentioned for the combinatorial version with analytic results. The following result is taken from [53, Proposition 6.4, p. 149].

LEMMA 9.13. *Let X be a simply-connected Riemannian manifold and $f : X \to \mathbb{R}$ be a function which is invariant under the isometries of X. Then there is a constant $C(f)$ with the property that for any cocompact free proper action of a discrete group Γ by isometries and any fundamental domain $\mathcal{F}$*

$$\int_{\mathcal{F}} f\, d\mathrm{vol}_X = C(f) \cdot \mathrm{vol}(X/\Gamma)$$

holds.

The next result is a consequence of the analytic definitions of L^2-Betti numbers, Novikov–Shubin invariants and L^2-torsion and Lemma 9.13.

LEMMA 9.14. *Let X be a complete Riemannian manifold. Then there are constants $B_p^{(2)}(X)$ for $p \geqslant 0$, $A_p(X)$ for $p \geqslant 1$ and $T(X)$ such that for any closed Riemannian* *manifold M whose universal covering is isometrically diffeomorphic to X the following holds*

$$\begin{aligned} b_p^{(2)}(M) &= B_p^{(2)}(X) \cdot \mathrm{vol}(M); \\ \alpha_p^{(2)}(M) &= A_p^{(2)}(X); \\ \rho^{(2)}(M) &= T^{(2)}(M) \cdot \mathrm{vol}(M). \end{aligned}$$

The constant $T^{(2)}(X)$ appearing in Lemma 9.14 can be computed for $X = \mathbb{H}^d$ as follows. Consider the polynomial with integer coefficients for $j \in \{0, 1, 2, \dots, n-1\}$

$$P_j^n(\nu) := \frac{\prod_{i=0}^n (\nu^2 + i^2)}{\nu^2 + (n-j)^2} = \sum_{k \geqslant 0}^{2n} K_{k,j}^n \cdot \nu^{2k}.$$

Define

$$\begin{aligned} C_d :=& \sum_{j=0}^{n-1} (-1)^{n+j+1} \frac{n!}{(2n)! \cdot \pi^n} \cdot \binom{2n}{j} \\ & \times \sum_{k=0}^{n} K_{k,j}^n \cdot \frac{(-1)^{k+1}}{2k+1} \cdot (n-j)^{2k+1}. \end{aligned} \tag{9.15}$$

The first values of C_d are computed in [118, Theorem 2]

$$\begin{aligned} C_3 &= \frac{1}{3\pi} \approx 0.106103; \\ C_5 &= \frac{62}{45\pi^2} \approx 0.139598; \\ C_7 &= \frac{221}{35\pi^3} \approx 0.203645; \\ C_{39} &\approx 4.80523 \cdot 10^7, \end{aligned}$$

and the constants C_d are positive, strictly increasing and grow very fast, namely they satisfy [118, Proposition 6]

$$\begin{aligned} C_{2n+1} &\geqslant \frac{n}{2\pi} \cdot C_{2n-1}; \\ C_{2n+1} &\geqslant \frac{2n!}{3(2\pi)^n}. \end{aligned}$$

The next result has been proven for 3-manifolds by Lott [142, Proposition 16] and Matthai [170, Corollary 6.7].

THEOREM 9.16 (Hess and Schick [118, Theorem 2]). *Let M be a closed hyperbolic d-dimensional manifold for odd $d = 2n+1$. Let $C_d > 0$ be the constant introduced in* (9.15). *Then*

$$\rho^{(2)}(\widetilde{M}) = (-1)^n \cdot C_d \cdot \mathrm{vol}(M).$$

Recall that we have introduced the basic notions and results about 3-manifolds in Section 3.

THEOREM 9.17 (Lück and Schick [164, Theorem 0.7]). *Let M be a compact connected orientable prime 3-manifold with infinite fundamental group such that the boundary of M is empty or a disjoint union of incompressible tori. Suppose that M satisfies Thurston's Geometrization Conjecture which implies that there is a decomposition along disjoint incompressible 2-sided tori in M whose pieces are Seifert manifolds or hyperbolic manifolds. Let $M_1, M_2, \dots, M_r$ be the hyperbolic pieces. They all have finite volume* [179, Theorem B, p. 52]. *Then M is admissible and*

$$\rho^{(2)}(\widetilde{M}) = -\frac{1}{3\pi} \cdot \sum_{i=1}^{r} \mathrm{vol}(M_i).$$

In particular, $\rho^{(2)}(\widetilde{M})$ is 0 if and and only if there are no hyperbolic pieces.

Examples of manifolds satisfying the assumptions of Theorem 9.17 are complements of knots in S^3.

We recall the definition of simplicial volume of an n-dimensional oriented closed manifold M [106, Section 0.2]. Let $C_*^{\mathrm{sing}}(M, \mathbb{R})$ be the singular chain complex of M with coefficients in the real numbers $\mathbb{R}$. An element c in $C_p^{\mathrm{sing}}(M, \mathbb{R})$ is given by a finite $\mathbb{R}$-linear combination $c = \sum_{i=1}^{s} r_i \cdot \sigma_i$ of singular p-simplices σ_i in M. Define the *l^1-norm* of c by setting

$$\|c\|_1 = \sum_{i=1}^{s} |r_i|.$$

For $\alpha \in H_m(M; \mathbb{R})$ define

$$\|\alpha\|_1 \;=\; \inf\bigl\{\|c\|_1 \mid c \in C_m^{\mathrm{sing}}(M; \mathbb{R}) \text{ is a cycle representing } \alpha\bigr\}.$$

The *simplicial volume* of M is defined by

$$\|M\| \;=\; \bigl\|[M]\bigr\|_1,$$

where $[M]$ is the image of the fundamental class of M under the change of ring homomorphism on singular homology $H_n(M; \mathbb{Z}) \to H_n(M; \mathbb{R})$. We mention the following extension of a question of Gromov [109, Section 8A].

CONJECTURE 9.18. Let M be a closed aspherical orientable manifold with vanishing simplicial volume. Then M is admissible and its L^2-torsion is trivial.

This conjecture is based on a variety of calculations and similarities of the properties of L^2-torsion and simplicial volume. The simplicial volume and the L^2-torsion are multiplicative under finite coverings (see [106, p. 8] and Theorem 9.9.7). If the universal coverings of two closed Riemannian manifolds M and N are isometrically diffeomorphic, then [106, p. 11]

$$\frac{\|M\|}{\operatorname{vol}(M)} = \frac{\|N\|}{\operatorname{vol}(N)}. \tag{9.19}$$

The corresponding proportionality principle holds for the L^2-torsion by Lemma 9.14. There are constants T_n and V_n, depending only on the dimension n, such that for any closed orientable hyperbolic manifold of dimension n we have

$$\|M\| = V_n \cdot \operatorname{vol}(M); \tag{9.20}$$
$$\rho^{(2)}(\widetilde{M}) = T_n \cdot \operatorname{vol}(M), \tag{9.21}$$

where V_n^{-1} is the supremum of all n-dimensional geodesic simplices, i.e., the convex hull of $(n+1)$ points in general position, in the n-dimensional hyperbolic space $\mathbb{H}^n$, and T_n is zero for even n and $T_3 = -\frac{1}{3\pi}$ (see [106, Section 2.2] and Theorem 9.16). Conjecture 9.18 is true for a closed aspherical orientable manifold with non-trivial smooth S^1-action. Namely, then its simplicial volume vanishes [106, Section 3.1], [253] and the map induced by evaluation $\pi_1(S^1) \to \pi_1(M)$ is injective [66, Lemma 5.1, p. 242 and Corollary 5.3, p. 243] which implies that M is admissible and $\rho^{(2)}(\widetilde{M}) = 0$ by Theorem 9.9.8. Theorem 9.17 is true for $\|M\|$ instead of $\rho^{(2)}(\widetilde{M})$, if one substitutes V_3 for $-\frac{1}{3\pi}$, by [236, 240]. Hence Conjecture 9.18 is true for 3-manifolds satisfying the hypothesis of Theorem 9.17.

The considerations above may suggest that one should conjecture for a closed oriented n-dimensional aspherical manifold M of dimension n that

$$\rho^{(2)}(\widetilde{M}) = C_n \cdot \|M\| \tag{9.22}$$

holds for a dimension constant C_n. This is true with $C_n = 0$ for even n and possibly true for $n = 3$ with $C_3 = -\frac{1}{3\pi \cdot V_3}$. However, it is definitely false in all odd dimensions $n \geqslant 9$ by the following argument. Let M be a 3-dimensional oriented closed hyperbolic 3-manifold. Let N be $M \times M \times M$ and F be an oriented closed hyperbolic surface. Let F^d be the d-fold Cartesian product of F with itself. We get from (9.20), (9.21), Theorem 9.9.5 and the product inequality for the simplicial volume in [106, p. 10]

$$\rho^{(2)}(\widetilde{N \times F^d}) = 0;$$
$$\rho^{(2)}(\widetilde{M \times F^{d+3}}) \neq 0;$$
$$\|N \times F^d\| \neq 0.$$

Since all the manifolds appearing in the list above are orientable closed aspherical manifolds of dimension $9 + 2d$, Equation (9.22) is wrong for odd $n \geqslant 9$.

The simplicial volume of an n-dimensional closed oriented manifold and its vanishing can be interpreted cohomologically as follows. Recall that a singular p-cochain on M with coefficients in $\mathbb{R}$ can be interpreted as a function $f : S_p(M) \to \mathbb{R}$, where $S_p(M)$ is the set of all singular p-simplices on M. Define its norm

$$\|f\|_\infty = \sup\{|f(s)| \mid s \in S_p(M)\}.$$

The *bounded cochain complex* $\widehat{C}^*(M)$ is the subcochain complex of the singular cochain complex $C^*(M;\mathbb{R})$ with real coefficients which consists of bounded cochains f, i.e., cochains f with $\|f\|_\infty < \infty$. The *bounded cohomology* $\widehat{H}^*(M)$ is the cohomology of this chain complex.

For $\beta \in H^n(M;\mathbb{R})$ define

$$\|\beta\|_\infty = \inf\{\|f\|_\infty \mid f \in C^n(M;\mathbb{R}) \text{ is a cocycle representing } \beta\} \in [0,\infty].$$

Let $\beta(M) \in H^n(M;\mathbb{R})$ be the cohomological fundamental class of M. Then we get

$$\begin{array}{ll} \|M\| = 0 & \text{if } \|\beta(M)\|_\infty = \infty; \\ \|M\| = \|\beta(M)\|_\infty^{-1} & \text{if } \|\beta(M)\|_\infty < \infty. \end{array}$$

Moreover, $\|M\|$ vanishes if and only if the canonical map $\widehat{H}^n(M) \to H^n(M;\mathbb{R})$ is trivial [17, pp. 278, 279], [106, p. 17].

If $f : M \to B\pi_1(M)$ is the classifying map of M, then the induced map on bounded cohomology $f^* : \widehat{H}^p(B\pi_1(M)) \to \widehat{H}^p(M)$ is an isometric isomorphism [106, p. 40], [126, p. 1105]. This implies that $\|M\|$ depends only on the image of the fundamental class under the classifying map $f_*([M]) \in H_n(B\pi_1(M);\mathbb{R})$. Namely, it is given by

$$\|M\| = \|f_*(\beta(M))\|_1. \tag{9.23}$$

Of course the L^2-torsion of a closed Riemannian manifold depends in general on more than $f_*([M]) \in H_n(B\pi_1(M);\mathbb{R})$, so that we see that Conjecture 9.18 only has a chance to be true for aspherical manifolds. Since the bounded cohomology $\widehat{H}^p(B\Gamma)$ vanishes for $p \geqslant 1$ for an amenable group, we get for any closed orientable manifold with amenable fundamental group that $\|M\| = 0$ [106, p. 40], [126, Theorem 4.3, p. 1105]. Hence Conjecture 9.18 implies for an aspherical closed Riemannian manifold with amenable fundamental group that its L^2-Betti numbers vanish and its L^2-torsion is trivial. We have already proven the statement for the L^2-Betti numbers in Theorem 4.1. We conjecture

CONJECTURE 9.24. If X is an aspherical connected finite CW-complex such that its fundamental group contains a non-trivial normal amenable subgroup, then its universal covering $\widetilde{X}$ is admissible and

$$\rho^{(2)}(\widetilde{X}) = 0.$$

Any closed smooth manifold satisfies

$$\|M\| \leqslant (n-1)^n n! \cdot \min \mathrm{vol}(M), \tag{9.25}$$

where $\min \mathrm{vol}(M)$ is the minimum over all volumes of M for all Riemannian metrics on M whose sectional curvature is pinched between -1 and 1 [106, p. 12]. So the vanishing of the minimal volume implies the vanishing of the simplicial volume. In view of Conjecture 9.18, the question arises whether for a closed aspherical manifold the vanishing of the minimal volume implies that M is admissible and its L^2-torsion is trivial. Under certain conditions on the sectional curvature there are also estimates of the volume from above by the simplicial volume due to Thurston [106, p. 10], [240].

Since the simplicial volume of a closed orientable manifold with non-trivial S^1-action vanishes [106, Section 3.1], [253], one could ask whether the vanishing of the simplicial volume is an obstruction to the existence of an S^1-foliation. Boileau, Druck and Vogt have dealt with this question in [27,28]. A positive answer would rule out the existence of an S^1-foliation on a closed hyperbolic manifold because the simplicial volume in that case is (up to a non-zero factor) the volume. In view of Conjecture 9.18, one could ask whether the existence of an S^1-foliation on a closed aspherical manifold implies the vanishing of all the L^2-Betti numbers and of the L^2-torsion. Again a positive answer to this question would settle the problem of the existence of an S^1-foliation on a closed hyperbolic manifold.

Suppose that the closed oriented manifold M admits a selfmap of degree different from 0 and ± 1. Then the simplicial volume is trivial [106, p. 8]. Assume additionally that M is aspherical. Then Conjecture 9.18 implies that $\rho^{(2)}(\widetilde{M})$ is trivial. This would be obvious if M would cover itself non-trivially. This raises the question of when a map $f: M \to M$ for M a closed aspherical manifold of positive degree d is homotopic to a covering of degree d.

Obviously the L^2-torsion is hard to compute, even in the combinatorial version where one does not have to deal with the regularization process. The problem is that it is very difficult to compute the spectral density function. Next we state a more algorithmic approach which was developed in [154]. Here we again exploit the fact that the combinatorial Laplace operator already lives over the integral group ring of the fundamental group.

Let $A \in M(n, m, \mathbb{C}\Gamma)$ be an (n, m)-matrix over $\mathbb{C}\Gamma$. It induces, by right multiplication, a $\mathbb{C}\Gamma$-homomorphism of left $\mathbb{C}\Gamma$-modules

$$R_A : \bigoplus_{i=1}^{n} \mathbb{C}\Gamma \to \bigoplus_{i=1}^{m} \mathbb{C}\Gamma, \qquad x \mapsto xA,$$

and by completion a bounded Γ-equivariant operator

$$R_A^{(2)} : \bigoplus_{i=1}^{n} l^2(\Gamma) \to \bigoplus_{i=1}^{m} l^2(\Gamma).$$

We define an involution of rings on $\mathbb{C}\Gamma$ by

$$\overline{\sum_{w\in\Gamma}\lambda_w\cdot w}=\sum_{w\in\Gamma}\overline{\lambda_w}\cdot w^{-1}.$$

Denote by A^* the (m,n)-matrix obtained from A by transposing and applying the involution above to each entry. As the notation suggests, the bounded Γ-equivariant operator $R_{A^*}^{(2)}$ is the adjoint of the bounded Γ-equivariant operator $R_A^{(2)}$. Define the *$\mathbb{C}\Gamma$-trace* of an element $u=\sum_{w\in\Gamma}\lambda_w\cdot w\in\mathbb{C}\Gamma$ by

$$\mathrm{tr}_{\mathbb{C}\Gamma}(u)=\lambda_e\in\mathbb{C}$$

for e the unit element in Γ. This extends to a square (n,n)-matrix A over $\mathbb{C}\Gamma$ by

$$\mathrm{tr}_{\mathbb{C}\Gamma}(A)=\sum_{i=1}^n\mathrm{tr}_{\mathbb{C}\Gamma}(a_{i,i})\in\mathbb{C}. \tag{9.26}$$

It follows directly from the definitions that the $\mathbb{C}\Gamma$-trace $\mathrm{tr}_{\mathbb{C}\Gamma}(A)$ of (9.26) agrees with the von Neumann trace $\mathrm{tr}_{\mathcal{N}(\Gamma)}(R_A^{(2)})$ defined in (1.2).

Let $A\in M(n,m,\mathbb{C}\Gamma)$ be an (n,m)-matrix over $\mathbb{C}\Gamma$. In the sequel let K be any positive real number satisfying

$$K\geqslant\|R_A^{(2)}\|$$

where $\|R_A^{(2)}\|$ is the operator norm of $R_A^{(2)}$. For $u=\sum_{w\in\Gamma}\lambda_w\cdot w\in\mathbb{C}\Gamma$, define $|u|_1=\sum_{w\in\Gamma}|\lambda_w|$. Then a possible choice for K is given by

$$K=\sqrt{m}\cdot\max\{|a_{i,j}|_1\mid 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant m\}.$$

The bounded Γ-equivariant operator

$$1-K^{-2}\cdot R_A^*R_A:\bigoplus_{i=1}^n l^2(\Gamma)\to\bigoplus_{i=1}^n l^2(\Gamma)$$

is positive. Let $(1-K^{-2}\cdot A^*A)^p$ be the p-fold product of matrices and $(1-K^{-2}\cdot R_A^*R_A)^p$ be the p-fold composition of operators.

DEFINITION 9.27. The *characteristic sequence* of a matrix $A\in M(n,m,\mathbb{C}\Gamma)$ and a nonnegative real number K satisfying $K\geqslant\|R_A^{(2)}\|$ is the sequence of real numbers

$$c(A,K)_p=\mathrm{tr}_{\mathbb{C}\Gamma}\big((1-K^{-2}\cdot AA^*)^p\big)=\mathrm{tr}_{\mathcal{N}(\Gamma)}\big((1-K^{-2}\cdot(R_A^{(2)})^*R_A^{(2)})^p\big).$$

THEOREM 9.28 (Lück [154, Theorem 4.4, p. 100]). *Let $A \in M(n, m, \mathbb{C}\Gamma)$ be an (n, m)-matrix over $\mathbb{C}\Gamma$. Denote by F the spectral density function of $R_A^{(2)}$. Let K be a positive real number satisfying $K \geqslant \|R_A^{(2)}\|$. Then*

(1) *The characteristic sequence $c(A, K)_p$ is a monotone decreasing sequence of non-negative real numbers;*

(2) *We have*

$$\dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr) = F(0) = \lim_{p\to\infty} c(A, K)_p;$$

(3) *Define $\beta(A) \in \mathbb{R}^{\geqslant 0} \cup \{\infty\}$ by*

$$\beta(A) = \sup\bigl\{\beta \in \mathbb{R}^{\geqslant 0} \mid \lim_{p\to\infty} p^{\beta} \cdot \bigl(c(A, K)_p - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr) = 0\bigr\}.$$

Then we have

$$\beta(A) \geqslant \alpha\bigl(R_A^{(2)}\bigr) \quad \text{if } \alpha\bigl(R_A^{(2)}\bigr) \text{ is a real number;}$$
$$\beta(A) = \infty \quad \text{otherwise;}$$

where $\alpha(R_A^{(2)})$ is the Novikov–Shubin invariant of Definition 8.4;

(4) *Let K be any positive real number satisfying $K \geqslant \|R_A^{(2)}\|$. Then the sum of positive real numbers*

$$\sum_{p=1}^{\infty} \frac{1}{p} \cdot \bigl(c(A, K)_p - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr)$$

converges if and only if $R_A^{(2)}$ is of determinant class, i.e., the integral $\int_{0+}^{\infty} \ln(\lambda)\, \mathrm{d}F$ converges. If $R_A^{(2)}$ is of determinant class, then

$$\begin{aligned} 2 \cdot \ln\bigl(\det_{\mathcal{N}(\Gamma)}\bigl(R_A^{(2)}\bigr)\bigr) &= 2 \cdot \bigl(n - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr) \cdot \ln(K) \\ &\quad - \sum_{p=1}^{\infty} \frac{1}{p} \cdot \bigl(c(A, K)_p - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr); \end{aligned}$$

(5) *Suppose $\alpha(R_A^{(2)}) > 0$. Then $\det_{\mathcal{N}(\Gamma)}(R_A^{(2)})$ is a positive real number. Given a real number α satisfying $0 < \alpha < \alpha(R_A^{(2)})$, there is a real number C such that we have for all $L \geqslant 1$*

$$0 \leqslant c(A, K)_L - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr) \leqslant \frac{C}{L^{\alpha}}$$

and

$$0 \leqslant -2 \cdot \ln\bigl(\det_{\mathcal{N}(\Gamma)}\bigl(R_A^{(2)}\bigr)\bigr) + 2 \cdot \bigl(n - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr) \cdot \ln(K)$$
$$- \sum_{p=1}^{L} \frac{1}{p} \cdot \bigl(c(A,K)_p - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr)\bigr) \leqslant \frac{C}{L^\alpha}.$$

Theorem 9.28 gives the possibility of computing

$$\dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(R_A^{(2)}\bigr)\bigr) \quad \text{and} \quad \ln\bigl(\det_{\mathcal{N}(\Gamma)}\bigl(R_A^{(2)}\bigr)\bigr)$$

by a sequence whose individual terms can be computed by an algorithm, provided a concrete presentation of Γ is given and the word problem can be solved for Γ. The speed of convergence can be predicted by the Novikov–Shubin invariants. However, we do not have a concrete value for the constant C appearing in Theorem 9.28.5. At any rate one gets upper bounds for $\dim_{\mathcal{N}(\Gamma)}(\ker(R_A^{(2)}))$ and $\ln(\det_{\mathcal{N}(\Gamma)}(R_A^{(2)}))$ since the characteristic sequence is monotone decreasing and positive. In this context Conjecture 2.1 is interesting. If, for instance, Γ is torsionfree and one of the elements of the characteristic sequence is smaller than 1, then Conjecture 2.1 implies that $\ker(R_A^{(2)})$ is trivial.

In particular, 3-manifolds are interesting since the cellular $\mathbb{Z}\Gamma$-chain complex of the universal covering can be computed from an appropriate presentation of the fundamental group. Theorem 9.28 implies [154, Theorem 2.4, p. 84]:

THEOREM 9.29. *Let M be a compact connected orientable irreducible 3-manifold with infinite fundamental group Γ. Let*

$$\Gamma = \langle s_1, s_2, \ldots, s_g \mid R_1, R_2, \ldots, R_r \rangle$$

be a presentation of Γ. Let the (r,g)-matrix

$$F = \begin{pmatrix} \frac{\partial R_1}{\partial s_1} & \cdots & \frac{\partial R_1}{\partial s_g} \\ \vdots & \ddots & \vdots \\ \frac{\partial R_r}{\partial s_1} & \cdots & \frac{\partial R_r}{\partial s_g} \end{pmatrix}$$

be the Fox matrix of the presentation. Denote by $\alpha_2(M)$ the second Novikov–Shubin invariant of M. Now there are two cases:

(1) *Suppose ∂M is non-empty. We make the assumption that ∂M is a union of incompressible tori and that $g = r - 1$. Then M is admissible. Define A to be the $(g-1, g-1)$-matrix with entries in $\mathbb{Z}\Gamma$ obtained from the Fox matrix by deleting one of the columns. Let α be any real number satisfying $0 < \alpha < \frac{2\cdot\alpha_2(M)}{\alpha_2(M)+2}$;*
(2) *Suppose ∂M is empty. We make the assumption that a finite covering of M is homotopy equivalent to a hyperbolic, Seifert or Haken 3-manifold and that the given presentation comes from a Heegaard decomposition. Then M is admissible and $g = r$. Define A to be the $(g-1, g-1)$-matrix with entries in $\mathbb{Z}\Gamma$ obtained from the Fox matrix by deleting one of the columns and one of the rows. Let α be any real number satisfying $0 < \alpha < \frac{2\cdot\alpha_2(M)}{\alpha_2(M)+1}$;*

Let K be any positive real number satisfying $K \geqslant \|R_A^{(2)}\|$. A possible choice for K is the product of $(g-1)^2$ and the maximum over the word length of those relations R_i whose Fox derivatives appear in A.

Then the sum of non-negative rational numbers $\sum_{p=1}^{L} \frac{1}{p} \cdot \mathrm{tr}_{\mathbb{Z}\pi}((1-K^{-2}\cdot AA^*)^p)$ converges to the real number $\rho^{(2)}(\widetilde{M}) + 2(g-1)\cdot \ln(K)$. More precisely, there is a constant $C > 0$ such that we get for all $L \geqslant 1$

$$0 \leqslant \rho^{(2)}(\widetilde{M}) + 2(g-1)\cdot \ln(K) - \sum_{p=1}^{L} \frac{1}{p} \cdot \mathrm{tr}_{\mathbb{Z}\pi}\big((1-K^{-2}\cdot AA^*)^p\big) \leqslant \frac{C}{L^{\alpha}}.$$

REMARK 9.30. Let M be a closed n-dimensional hyperbolic manifold. Then Mostow's Rigidity Theorem says that the isometric diffeomorphism type, and in particular the volume of M, depends only on its fundamental group [17, Theorem C.0, p. 83], [182]. We get from Theorem 9.12, Theorem 9.16 and Theorem 9.29 a way of computing the volume purely in terms of a presentation of the fundamental group without using information about M itself. If for a group Γ the classifying space $B\Gamma$ is a finite CW-complex and $E\Gamma$ is admissible, then its L^2-torsion is defined and is a generalization of the volume in the case where Γ is the fundamental group of an odd-dimensional hyperbolic closed manifold.

EXAMPLE 9.31. In [154, Example 2.7] the complement M of the figure eight knot is computed. For the presentation of $\Gamma = \pi_1(M)$

$$\Gamma = \langle s_1, s_2, t \mid ts_1t^{-1}s_2^{-1} = ts_2t^{-1}s_1s_2^{-3} = 1\rangle$$

and the $(2,2)$-matrix

$$B = \begin{pmatrix} 13 + s_2 + s_2^{-1} & -1 + s_2 + s_1s_2^3 - s_2s_1s_2^{-3} - ts_1s_2^{-3} \\ -1 + s_2^{-1} + s_2^3s_1^{-1} - s_2^3s_1^{-1}s_2^{-1} - s_2^3s_1^{-1}t^{-1} & 13 + s_2^3s_1^{-1} + s_1s_2^{-3} \end{pmatrix}$$

we obtain

$$\rho^{(2)}(\widetilde{M}) = -8\ln(2) + \sum_{p=1}^{\infty} \frac{1}{p\cdot 16^p} \cdot \mathrm{tr}_{\mathbb{Z}\Gamma}(B^p).$$

We have already mentioned in Conjecture 2.1 and Conjecture 8.9 what possible values we expect for the L^2-Betti numbers and Novikov–Shubin invariants. We do not have a good guess in the case of L^2-torsion for spaces of acyclic determinant class. This question only makes sense if the space is L^2-acyclic, otherwise one could vary the Riemannian metric to get any real number as the L^2-torsion. Recall that for an (m,n)-matrix A, we denote by $R_A : l^2(\Gamma)^m \to l^2(\Gamma)^n$ the bounded Γ-equivariant operator induced by right multiplication with A.

DEFINITION 9.32. For a group Γ, define a multiplicative subgroup of the positive real numbers

$$\begin{aligned} R(\Gamma) = &\{\det_{\mathcal{N}(\Gamma)}(R_A) \mid A \in M(n,n,\mathbb{Z}\Gamma),\ \det_{\mathcal{N}(\Gamma)}(R_A) \neq 0,\ \ker(R_A) = 0\} \\ &\cup \{(\det_{\mathcal{N}(\Gamma)}(R_A))^{-1} \mid A \in M(n,n,\mathbb{Z}\Gamma),\ \det_{\mathcal{N}(\Gamma)}(R_A) \neq 0, \\ &\qquad \ker(R_A) = 0\}. \end{aligned}$$

LEMMA 9.33. *Let Γ be a finitely presented group Γ. Suppose that there is at least one connected finite CW-complex Y with $\Gamma = \pi_1(Y)$ such that $\widetilde{Y}$ is of acyclic determinant class. Then*

$$\begin{aligned} 2 \cdot \ln(R(\Gamma)) &\subset \{\rho^{(2)}(\widetilde{X}) \mid X \text{ a connected finite } CW\text{-complex with } \pi_1(X) = \Gamma \\ &\qquad \text{and } \widetilde{X} \text{ of acyclic determinant class}\} \\ &\subset \ln(R(\Gamma)) \end{aligned}$$

and

$$\begin{aligned} 4 \cdot \ln(R(\Gamma)) &\subset \{\rho^{(2)}(\widetilde{M}) \mid M \text{ a closed manifold with } \pi_1(M) = \Gamma \\ &\qquad \text{and } \widetilde{M} \text{ of acyclic determinant class}\} \\ &\subset \ln(R(\Gamma)). \end{aligned}$$

PROOF. The non-trivial inclusion in the first assertion is

$$\begin{aligned} 2 \cdot \ln(R(\Gamma)) \subset \{\rho^{(2)}(\widetilde{X}) \mid\ & X \text{ a connected finite } CW\text{-complex with } \pi_1(X) = \Gamma \\ & \text{and } \widetilde{X} \text{ of acyclic determinant class}\}. \end{aligned}$$

Let $Z = Y \times S^3$. Then $\pi_1(Z) = \Gamma$ and $\widetilde{Z}$ is of acyclic determinant class with $\rho^{(2)}(\widetilde{Z}) = 0$ by Theorem 9.9.5. Let A be an (n,n)-matrix over $\mathbb{Z}\Gamma$ such that $\ker(R_A) = 0$ and $\det_{\mathcal{N}(\Gamma)}(R_A) \neq 0$. Let n be an integer such that $2n$ is greater than or equal to the dimension of Z. By attaching cells to Z in dimensions $2n+2$ and $2n+3$ we obtain a connected finite CW-complex X such that $\Gamma = \pi_1(X)$ and the cellular $\mathbb{Z}\Gamma$-chain complex of $\widetilde{X}$ is the direct sum of the one of $\widetilde{Z}$ and the chain complex concentrated in dimensions $2n+2$ and $2n+3$ whose only non-trivial differential is given by R_A. Then Lemma 9.5 implies that $\widetilde{X}$ is of determinant class and

$$\begin{aligned} \rho^{(2)}(\widetilde{X}) &= -(-1)^{2n+3} \cdot (n+3) \cdot \ln(\det_{\mathcal{N}(\Gamma)}(R_A^* R_A)) \\ &\quad - (-1)^{2n+2} \cdot (n+2) \cdot \ln(\det_{\mathcal{N}(\Gamma)}(R_A R_A^*)) \\ &= 2 \cdot \ln(\det_{\mathcal{N}(\Gamma)}(R_A)). \end{aligned}$$

This shows the first assertion.

Let X be a connected finite CW-complex such that $\Gamma = \pi_1(X)$ and $\widetilde{X}$ is of acyclic determinant class. Let n be an integer such that $2n$ is greater or equal to the dimension of X. We embed X into $\mathbb{R}^{2n+2}$. Let M be the boundary of a regular neighborhood N

of X. Now L^2-torsion and the notions of acyclic determinant class can also be defined for pairs and simple homotopy invariance, the sum formula and Poincaré duality extend to this case [154]. Since the inclusion of X into N is a simple homotopy equivalence and X is of acyclic determinant class, $\widetilde{N}$ by homotopy invariance, $(\widetilde{N}, \widetilde{M})$ by Poincaré duality, and hence $\widetilde{M}$ by additivity (= formula for pairs) are of acyclic determinant class, and we get

$$\begin{aligned} \rho^{(2)}(\widetilde{M}) &= \rho^{(2)}(\widetilde{N}) - \rho^{(2)}(\widetilde{N}, \widetilde{M}) = \rho^{(2)}(\widetilde{N}) + \rho^{(2)}(\widetilde{N}) \\ &= 2 \cdot \rho^{(2)}(\widetilde{N}) = 2 \cdot \rho^{(2)}(\widetilde{X}). \end{aligned}$$

This finishes the proof of Lemma 9.33. □

If Γ is countable, then $R(\Gamma)$ is countable because then $\mathbb{Z}\Gamma$, and hence $M(n, n, \mathbb{Z}\Gamma)$, is countable for all n and $R(\Gamma)$ is a countable union of countably sets.

If $\Gamma = \mathbb{Z}$, then each element in $R(\Gamma)$ is an algebraic number, i.e., the root of a non-trivial polynomial with rational coefficients, by the following argument.

We get from [163, Section 4]

$$R(\mathbb{Z}) = \{\det_{\mathcal{N}(\mathbb{Z})}(R_p) \mid p \in \mathbb{Z}[\mathbb{Z}],\ p \neq 0\}.$$

We can write p in $\mathbb{C}[\mathbb{Z}]$ as

$$p(z) = C \cdot z^n \cdot \prod_{k=1}^{l} (z - a_i)$$

for complex numbers $C, a_0, a_1, \ldots, a_l$ and integers n and $l \geqslant 0$. Since p is a non-zero polynomial with integer coefficients, C must be a non-zero integer and each a_i is algebraic. We get from Lemma 9.5.4

$$\begin{aligned} \det_{\mathcal{N}(\mathbb{Z})}(R_p) &= \det_{\mathcal{N}(\mathbb{Z})}(R_C) \cdot \det_{\mathcal{N}(\mathbb{Z})}(R_{z^n}) \cdot \prod_{k=1}^{l} \det_{\mathcal{N}(\mathbb{Z})}(R_{(z-a_i)}) \\ &= \prod_{k=1}^{l} \det_{\mathcal{N}(\mathbb{Z})}(R_{(z-a_i)}). \end{aligned}$$

Hence the claim follows from the following equation

$$\det_{\mathcal{N}(\mathbb{Z})}(R_{(z-a)}) = \begin{cases} |a| & \text{for } |a| \geqslant 1, \\ 1 & \text{for } |a| \leqslant 1, \end{cases} \tag{9.34}$$

which we prove next. We have to show that, for $a \in \mathbb{C}$, if we equip S^1 with the obvious measure satisfying $\operatorname{vol}(S^1) = 1$

$$\int_{S^1} \ln\big((z-a)(z^{-1} - \bar{a})\big)\, d\operatorname{vol} = \begin{cases} 2 \cdot \ln(|a|) & \text{for } |a| \geqslant 1, \\ 0 & \text{for } |a| \leqslant 1. \end{cases} \tag{9.35}$$

We have

$$\int_{S^1} \ln\big((z-a)(z^{-1}-\overline{a})\big)\,\mathrm{d}\,\mathrm{vol} = \int_{S^1} \ln\big((z-|a|)(z^{-1}-|a|)\big)\,\mathrm{d}\,\mathrm{vol}.$$

Hence we may suppose in the sequel $a \in \mathbb{R}^{\geqslant 0}$.

We compute for $a \neq 1$ and the path $\gamma : [0,1] \to S^1$, $t \mapsto \exp(2\pi i t)$, using the Residue Theorem

$$\begin{aligned}
&\int_{S^1} \frac{\mathrm{d}}{\mathrm{d}a} \ln\big((z-a)(z^{-1}-a)\big)\,\mathrm{d}\,\mathrm{vol} = \int_{S^1} \frac{1}{a-z} + \frac{1}{a-z^{-1}}\,\mathrm{d}\,\mathrm{vol} \\
&= 2 \cdot \int_{S^1} \frac{1}{a-z}\,\mathrm{d}\,\mathrm{vol} = 2 \cdot \int_{S^1} \frac{1}{(a-z)\cdot 2\pi \mathrm{i} z} \cdot 2\pi \mathrm{i} z \cdot \mathrm{d}\,\mathrm{vol} \\
&= \frac{2}{2\pi \mathrm{i}} \cdot \int_\gamma \frac{1}{(a-z)\cdot z}\,\mathrm{d}z \\
&= \begin{cases} \frac{2}{a} & \text{for } a > 1, \\ 0 & \text{for } a < 1. \end{cases}
\end{aligned}$$

This implies for $a \in \mathbb{R}^{\geqslant 0}$, $a \neq 1$

$$\frac{\mathrm{d}}{\mathrm{d}a} \int_{S^1} \ln\big((z-a)(z^{-1}-a)\big)\,\mathrm{d}\,\mathrm{vol} = \begin{cases} \frac{2}{a} & \text{for } a > 1, \\ 0 & \text{for } a < 1. \end{cases}$$

We conclude for an appropriate number C

$$\int_{S^1} \ln\big((z-a)(z^{-1}-a)\big)\,\mathrm{d}\,\mathrm{vol} = \begin{cases} 2 \cdot \ln(a) + C & \text{for } a > 1, \\ 0 & \text{for } a < 1. \end{cases}$$

We get from Levi's Theorem of Monotone Convergence

$$\int_{S^1} \ln\big((z-1)(z^{-1}-1)\big)\,\mathrm{d}\,\mathrm{vol} = C.$$

We get from Lebesgue's Theorem of Majorized Convergence

$$\int_{S^1} \ln\big((z-1)(z^{-1}-1)\big)\,\mathrm{d}\,\mathrm{vol} = 0.$$

This proves (9.35), and hence (9.34).

More information about L^2-torsion can be found in [42,44,47,48,70,71,146,172].

10. Algebraic dimension theory of finite von Neumann algebras

In this section we give a purely algebraic approach to L^2-Betti numbers and the von Neumann dimension and extend all these notions for a finitely generated Hilbert $\mathcal{N}(\Gamma)$-module which is essentially the same as a finitely generated projective $\mathcal{N}(\Gamma)$-module to arbitrary $\mathcal{N}(\Gamma)$-modules.

We mention the following observations and facts which first were made by Farber [89,90] and then independently by the author [157]. Farber's approach is different in that he works with a more abstract category than the category of finitely presented modules. Both approaches are compared and identified in [157, Theorem 0.9].

There is an equivalence of categories

$$\nu : \{\text{fin. gen. proj.}\ \mathcal{N}(\Gamma)\text{-mod.}\} \to \{\text{fin. gen. Hilb.}\ \mathcal{N}(\Gamma)\text{-mod.}\} \tag{10.1}$$

where {fin. gen. proj. $\mathcal{N}(\Gamma)$-mod.} is the category of finitely generated projective modules over the ring $\mathcal{N}(\Gamma)$ with $\mathcal{N}(\Gamma)$-linear maps as morphisms, and {fin. gen. Hilb. $\mathcal{N}(\Gamma)$-mod.} is the category of finitely generated Hilbert $\mathcal{N}(\Gamma)$-modules with bounded $\mathcal{N}(\Gamma)$-equivariant operators as morphisms [157, Theorem 2.2]. It sends $\mathcal{N}(\Gamma)$ to $l^2(\Gamma)$. It is compatible with finite direct sums, with the complex vector space structures on the set of morphisms and with the involutions given by taking dual $\mathcal{N}(\Gamma)$-modules and dual homomorphisms in {fin. gen. proj. $\mathcal{N}(\Gamma)$-mod.} and adjoint operators in {fin. gen. Hilb. $\mathcal{N}(\Gamma)$-mod.}. The category of finitely generated projective $\mathcal{N}(\Gamma)$-modules is a subcategory of the category of finitely presented $\mathcal{N}(\Gamma)$-modules

$$\{\text{fin. gen. proj.}\ \mathcal{N}(\Gamma)\text{-mod.}\} \subset \{\text{fin. pres.}\ \mathcal{N}(\Gamma)\text{-mod.}\}. \tag{10.2}$$

The point is that $\mathcal{N}(\Gamma)$ is a semi-hereditary ring, i.e., finitely generated $\mathcal{N}(\Gamma)$-submodules of projective $\mathcal{N}(\Gamma)$-modules are projective and the category {fin. pres. $\mathcal{N}(\Gamma)$-mod.} is abelian, i.e., the kernel, the image and the cokernel of an $\mathcal{N}(\Gamma)$-linear map of finitely presented $\mathcal{N}(\Gamma)$-modules are again finitely presented [90, §2], [157, Theorem 1.2 and Corollary 2.4]. Let M be an $\mathcal{N}(\Gamma)$-submodule of N. Define the *closure of M in N* to be the $\mathcal{N}(\Gamma)$-submodule of N

$$\begin{aligned}\overline{M} = \{x \in N \mid f(x) = 0 \text{ for all } f \in \hom_{\mathcal{N}(\Gamma)}(N, \mathcal{N}(\Gamma)) \\ \text{with } M \subset \ker(f)\}.\end{aligned} \tag{10.3}$$

The functor ν of (10.1) respects exact and weakly exact sequences [157, Lemma 2.3] where weakly exact for {fin. gen. Hilb. $\mathcal{N}(\Gamma)$-mod.} was defined in Section 1 and translates to {fin. gen. proj. $\mathcal{N}(\Gamma)$-mod.} using (10.3). For an $\mathcal{N}(\Gamma)$-module M define the $\mathcal{N}(\Gamma)$-submodule $\mathbf{T}M$ and the $\mathcal{N}(\Gamma)$-quotient module $\mathbf{P}M$ by (see also [90, §3])

$$\mathbf{T}M = \{x \in M \mid f(x) = 0 \text{ for all } f \in \hom_{\mathcal{N}(\Gamma)}(M, \mathcal{N}(\Gamma))\}; \tag{10.4}$$

$$\mathbf{P}M = M/\mathbf{T}M. \tag{10.5}$$

Notice that $\mathbf{T}M$ is the closure of the trivial module in M. If M is finitely generated, then $\mathbf{P}M$ is finitely generated projective [157, Theorem 1.2]. These notions now allow one to read off the L^2-Betti numbers and the Novikov–Shubin invariants of a regular covering $\overline{X} \to X$ with Γ as group of deck transformations from the finitely presented $\mathcal{N}(\Gamma)$-module

$H_p(\mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} C(\overline{X}))$ [90,157]. Moreover, one can generalize all these invariants using the universal center-valued trace of $\mathcal{N}(\Gamma)$ as carried out in [157].

Because of (10.1), one can define for a finitely generated projective $\mathcal{N}(\Gamma)$-module

$$\dim_{\mathcal{N}(\Gamma)}(P) = \dim_{\mathcal{N}(\Gamma)}\bigl(\nu(P)\bigr) \in [0, \infty) \tag{10.6}$$

where $\dim_{\mathcal{N}(\Gamma)}(\nu(P))$ is defined in Definition 1.3.

THEOREM 10.7 (Lück, [159, Theorem 0.6]). *There is a dimension function*

$$\dim_{\mathcal{N}(\Gamma)} : \{\mathcal{N}(\Gamma)\text{-modules}\} \to [0, \infty]$$

which has the following properties:

(1) *Extension property*
If M is finitely generated projective, then $\dim_{\mathcal{N}(\Gamma)}(M)$ *agrees with the number given in* (10.6);

(2) *Invariance under closure*
If $K \subset M$ is a submodule of the finitely generated $\mathcal{N}(\Gamma)$-module M, then

$$\dim_{\mathcal{N}(\Gamma)}(K) = \dim_{\mathcal{N}(\Gamma)}\bigl(\overline{K}\bigr);$$

(3) *Cofinality*
Let $\{M_i \mid i \in I\}$ be a cofinal system of submodules of M, i.e., $M = \bigcup_{i \in I} M_i$ and for two indices i and j there is an index k in I satisfying $M_i, M_j \subset M_k$. Then

$$\dim_{\mathcal{N}(\Gamma)}(M) = \sup\bigl\{\dim_{\mathcal{N}(\Gamma)}(M_i) \mid i \in I\bigr\};$$

(4) *Additivity*
If $0 \to M_0 \xrightarrow{i} M_1 \xrightarrow{p} M_2 \to 0$ is an exact sequence of $\mathcal{N}(\Gamma)$-modules, then

$$\dim_{\mathcal{N}(\Gamma)}(M_1) = \dim_{\mathcal{N}(\Gamma)}(M_0) + \dim_{\mathcal{N}(\Gamma)}(M_2);$$

(5) *If M is a finitely generated $\mathcal{N}(\Gamma)$-module, then*

$$\dim_{\mathcal{N}(\Gamma)}(M) = \dim_{\mathcal{N}(\Gamma)}(\mathbf{P}M);$$
$$\dim_{\mathcal{N}(\Gamma)}(\mathbf{T}M) = 0;$$

(6) *Uniqueness*
This dimension function is uniquely determined by the extension property, invariance under closure, cofinality and additivity.

Meanwhile this dimension function (and its properties) has been extended from the von Neumann algebra $\mathcal{N}(\Gamma)$ to the associated algebra of affiliated operators by Reich [213].

Let $i: \Delta \to \Gamma$ be an injective group homomorphism. We claim that associated to i there is a ring homomorphism of the group von Neumann algebras, also denoted by

$$i: \mathcal{N}(\Delta) \to \mathcal{N}(\Gamma).$$

Recall from (1.1) that $\mathcal{N}(\Delta)$ is the same as the ring $\mathcal{B}(l^2(\Delta), l^2(\Delta))^{\Delta}$ of bounded Δ-equivariant operators $f: l^2(\Delta) \to l^2(\Delta)$. Notice that $\mathbb{C}\Gamma \otimes_{\mathbb{C}\Delta} l^2(\Delta)$ can be viewed as a dense subspace of $l^2(\Gamma)$ and that f defines a $\mathbb{C}\Gamma$-homomorphism $\mathrm{id} \otimes_{\mathbb{C}\Delta} f: \mathbb{C}\Gamma \otimes_{\mathbb{C}\Delta} l^2(\Delta) \to \mathbb{C}\Gamma \otimes_{\mathbb{C}\Delta} l^2(\Delta)$ which is bounded with respect to the pre-Hilbert structure induced on $\mathbb{C}\Gamma \otimes_{\mathbb{C}\Delta} l^2(\Delta)$ from $l^2(\Gamma)$. Hence $\mathrm{id} \otimes_{\mathbb{C}\Delta} f$ extends to a Γ-equivariant bounded operator $i(f): l^2(\Gamma) \to l^2(\Gamma)$.

Given an $\mathcal{N}(\Delta)$-module M, define *induction with i* to be the $\mathcal{N}(\Gamma)$-module

$$i_*(M) = \mathcal{N}(\Gamma) \otimes_{\mathcal{N}(\Delta)} M.$$

Obviously i_* is a covariant functor from the category of $\mathcal{N}(\Delta)$-modules to the category of $\mathcal{N}(\Gamma)$-modules, preserves direct sums and the properties finitely generated and projective and sends $\mathcal{N}(\Delta)$ to $\mathcal{N}(\Gamma)$. We get from [159, Theorem 3.3]

THEOREM 10.8. *Let $i: \Delta \to \Gamma$ be an injective group homomorphism. Then*

(1) *i_* is an exact functor, i.e., for any exact sequence of $\mathcal{N}(\Delta)$-modules $M_0 \to M_1 \to M_2$ the induced sequence of $\mathcal{N}(\Gamma)$-modules $i_*M_0 \to i_*M_1 \to i_*M_2$ is exact;*

(2) *For any $\mathcal{N}(\Delta)$-module M we have*

$$\dim_{\mathcal{N}(\Delta)}(M) = \dim_{\mathcal{N}(\Gamma)}(i_*M).$$

This allows us to extend the definition of L^2-Betti numbers to arbitrary topological spaces with Γ-action and to arbitrary groups.

DEFINITION 10.9. Let Γ be a group acting on the topological space Z. Let $C_*^{\mathrm{sing}}(Z)$ be the singular chain complex, which becomes a $\mathbb{Z}\Gamma$-chain complex by the Γ-action. Define

$$H_p^{\Gamma}(Z; \mathcal{N}(\Gamma)) = H_p(\mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} C_*^{\mathrm{sing}}(Z));$$
$$b_p^{(2)}(Z) = \dim_{\mathcal{N}(\Gamma)}(H_p^{\Gamma}(Z; \mathcal{N}(\Gamma))) \in [0, \infty].$$

Given a group Γ, define

$$b_p^{(2)}(\Gamma) = b_p^{(2)}(E\Gamma) \in [0, \infty].$$

These L^2-Betti numbers are investigated in [159,160] and [213], where also their relation with the generalized L^2-Betti numbers defined in [54] is explained. Obviously they depend only on the Γ-homotopy type of Z and they agree with the L^2-Betti numbers of Definition 1.5 in the special case where Z is the total space $\overline{X}$ of a regular covering $\overline{X} \to X$ of a CW-complex X of finite type with Γ as deck transformation group. The point of this extension is that it is useful to have the notion of L^2-Betti numbers for arbitrary Γ-spaces, even if one wants to compute them only for a regular covering of a CW-complex of finite type. For instance in Theorem 4.1 one wants to compute the L^2-Betti numbers of $E\Gamma$ in the case where $B\Gamma$ is assumed to be of finite type, using the information that Γ contains a non-trivial amenable subgroup $\Delta \subset \Gamma$, but no information on $B\Delta$ is given. Next we sketch how Theorem 4.1 follows from Lemma 4.4. Details of this proof and a comparison with the original proof in [54] are given in [159, Section 5]. We will show the more general result.

THEOREM 10.10.

(1) *Let Γ be an infinite amenable group. Then*

$$b_p^{(2)}(\Gamma) = 0 \quad \textit{for } p \geqslant 0;$$

(2) *Let Δ be a normal subgroup of Γ with $b_p^{(2)}(\Delta) = 0$ for $p \geqslant 0$. Then*

$$b_p^{(2)}(\Gamma) = 0 \quad \textit{for } p \geqslant 0.$$

PROOF. (1) In the sequel colimits are taken over the directed system of finite subcomplexes Y of $B\Gamma$ and $\overline{Y}$ is the restriction of the universal Γ-principal bundle $E\Gamma \to B\Gamma$ to Y. Notice that a colimit over a directed system is an exact functor and compatible with tensor products. Hence the following diagram commutes and has isomorphisms as horizontal maps

$$\begin{array}{ccc} \operatorname{colim} \mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} H_p(\overline{Y}) & \xrightarrow{\cong} & \mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} H_p(E\Gamma) \\ {\scriptstyle i_1}\downarrow & & \downarrow{\scriptstyle i_2} \\ \operatorname{colim} H_p^{\Gamma}(\overline{Y}; \mathcal{N}(\Gamma)) & \xrightarrow[\cong]{} & H_p^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma)) \end{array}$$

where the horizontal arrows are given by the inclusions of $\overline{Y}$ into $E\Gamma$ and the vertical arrows are the canonical maps. It is not hard to deduce from Lemma 4.4 the in (some sense dual) statement that the dimension $\dim_{\mathcal{N}(\Gamma)}$ of the cokernel of each of the maps

$$\mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} H_p(\overline{Y}) \to H_p^{\Gamma}(\overline{Y}; \mathcal{N}(\Gamma))$$

is zero because Γ is amenable. Since colimit over a directed system is an exact functor, we conclude from cofinality and additivity of $\dim_{\mathcal{N}(\Gamma)}$ of Theorem 10.7 that the dimension $\dim_{\mathcal{N}(\Gamma)}$ of the cokernel of the left vertical arrow in the diagram above is zero. From additivity of the dimension $\dim_{\mathcal{N}(\Gamma)}$ (see Theorem 10.7) we conclude

$$\dim_{\mathcal{N}(\Gamma)}\big(\mathcal{N}(\Gamma) \otimes_{\mathbb{Z}\Gamma} H_p(E\Gamma)\big) \geqslant \dim_{\mathcal{N}(\Gamma)}\big(H_p^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma))\big).$$

Since $H_p(E\Gamma)$ vanishes for $p \geqslant 1$, we get $b_p^{(2)}(\Gamma) = 0$ for $p \geqslant 1$. Since Γ is infinite, a direct calculation shows $b_0^{(2)}(\Gamma) = 0$.

(2) We have the fibration $B\Delta \to B\Gamma \to B\pi$ for $\pi = \Gamma/\Delta$. Since $B\pi$ is a CW-complex and we are dealing with homology, the associated Leray–Serre spectral sequence with coefficients in $\mathcal{N}(\Gamma)$ converges to $H_p^\Gamma(E\Gamma; \mathcal{N}(\Gamma))$. Its E^2-term is given by

$$E^2_{p,q} = H_p^\pi\big(E\pi; H_q^\Gamma\big(\Gamma \times_\Delta E\Delta; \mathcal{N}(\Gamma)\big)\big)$$

for a certain action of π on the $\mathcal{N}(\Gamma)$-module $H_q^\Gamma(\Gamma \times_\Delta E\Delta; \mathcal{N}(\Gamma))$ which comes from the fiber transport [150, Section 1], [151, Section 4]. We get from Theorem 10.8 and the assumption

$$\begin{aligned}
&\dim_{\mathcal{N}(\Gamma)}\big(H_q^\Gamma\big(\Gamma \times_\Delta E\Delta; \mathcal{N}(\Gamma)\big)\big)\\
&\quad = \dim_{\mathcal{N}(\Delta)}\big(\mathcal{N}(\Gamma) \otimes_{\mathcal{N}(\Delta)} H_q^\Delta\big(E\Delta; \mathcal{N}(\Delta)\big)\big)\\
&\quad = \dim_{\mathcal{N}(\Delta)}\big(H_q^\Delta\big(E\Delta; \mathcal{N}(\Delta)\big)\big)\\
&\quad = 0.
\end{aligned}$$

Now we conclude from additivity of the dimension $\dim_{\mathcal{N}(\Gamma)}$ (see Theorem 10.7)

$$\begin{aligned}
&\dim_{\mathcal{N}(\Gamma)}\big(E^2_{p,q}\big) = 0 && \text{for } p, q \geqslant 0;\\
&\dim_{\mathcal{N}(\Gamma)}\big(H_p^\Gamma\big(E\Gamma; \mathcal{N}(\Gamma)\big)\big) = 0 && \text{for } p \geqslant 0.
\end{aligned}$$

This finishes the proof of Theorem 10.10. □

More information about extending the definition of L^2-cohomology can be found in [91, 235]. The same program has been carried out for Novikov–Shubin invariants in [162].

11. The zero-in-the-spectrum Conjecture

In this section we deal with the zero-in-the-spectrum Conjecture. To our knowledge this conjecture is due to John Lott. We recommend to the reader the survey article [144] where more information can be found. Gromov deals with this problem in the aspherical case in [109].

CONJECTURE 11.1 (*Zero-in-the-spectrum Conjecture*). There is no connected closed Riemannian manifold M such that for all $p \geqslant 0$ zero is not in the spectrum of the Laplace operator Δ_p acting on smooth p-forms of the universal covering $\widetilde{M}$.

Lott [144] gives five versions of this conjecture, stated as a question, namely, that for some $p \geqslant 0$ zero is in the spectrum of the Laplace operator Δ_p acting on smooth p-forms of $\widetilde{M}$ if

(1) $\widetilde{M}$ is a complete Riemannian manifold;
(2) $\widetilde{M}$ is a complete Riemannian manifold with bounded geometry, i.e., the injectivity radius is positive and the sectional curvature is pinched between -1 and 1;

(3) $\widetilde{M}$ is a uniformly contractible Riemannian manifold, i.e., for all $r > 0$ there is an $R(r) > 0$ such that for all $m \in M$ the metric ball $B_r(m)$ is contractible within $B_{R(r)}(m)$;
(4) $\widetilde{M}$ is the universal covering of a closed Riemannian manifold;
(5) $\widetilde{M}$ is the universal covering of a closed aspherical Riemannian manifold.

We emphasize that the next definition makes sense for arbitrary groups and spaces.

DEFINITION 11.2. Let Z be a topological space with an action of the group Γ. Let $H_p^\Gamma(Z; \mathcal{N}(\Gamma))$ be the $\mathcal{N}(\Gamma)$-module given by the singular homology with coefficients in $\mathcal{N}(\Gamma)$ as defined in 10.9. We say that Z is *$\mathcal{N}(\Gamma)$-acyclic* (respectively *n-$\mathcal{N}(\Gamma)$-connected*) if $H_p^\Gamma(Z; \mathcal{N}(\Gamma))$ vanishes for $p \geqslant 0$ (respectively $0 \leqslant p \leqslant n$).

A group Γ is called *$\mathcal{N}(\Gamma)$-acyclic* respectively *n-$\mathcal{N}(\Gamma)$-connected* if the universal Γ-space $E\Gamma$ has this property.

LEMMA 11.3. *The following statements are equivalent for a finitely presented group Γ:*
(1) *There is no connected closed Riemannian manifold M with fundamental group Γ such that, for all $p \geqslant 0$, zero is not in the spectrum of the Laplace operator Δ_p acting on smooth p-forms of the universal covering $\widetilde{M}$;*
(2) *There is no connected closed Riemannian manifold M with fundamental group Γ such that*

$$b_p^{(2)}(\widetilde{M}) = 0, \qquad p \geqslant 0;$$
$$\alpha_p(\widetilde{M}) = \infty^+, \qquad p \geqslant 0;$$

(3) *There is no connected finite CW-complex X with fundamental group Γ such that*

$$b_p^{(2)}(\widetilde{X}) = 0, \qquad p \geqslant 0;$$
$$\alpha_p(\widetilde{X}) = \infty^+, \qquad p \geqslant 0;$$

(4) *There is no connected finite CW-complex X with fundamental group Γ such that for all $p \geqslant 0$ the combinatorial Laplace operator Δ_p acting on the pth chain module of the L^2-chain complex of the universal covering $\widetilde{X}$ is invertible;*
(5) *There is no connected finite CW-complex X with fundamental group Γ such that $\widetilde{X}$ is $\mathcal{N}(\Gamma)$-acyclic.*

PROOF. The equivalence (1) $\Leftrightarrow$ (2) follows from the analytic definition of L^2-Betti numbers and Novikov–Shubin invariants.

(2) $\Leftrightarrow$ (3) The implication (3) $\Rightarrow$ (2) is obvious since a closed manifold is a finite CW-complex. To prove (2) $\Rightarrow$ (3), let X be a connected finite CW-complex with fundamental group Γ such that

$$b_p^{(2)}(\widetilde{X}) = 0, \qquad p \geqslant 0;$$
$$\alpha_p(\widetilde{X}) = \infty^+, \qquad p \geqslant 0;$$

Let n be the dimension of X. Since X cannot be S^1 because of Example 8.5, and cannot be $\bigvee_{i=1}^r S^1$ for $r \geqslant 2$ because the Euler characteristic of X must be trivial (see Theorem 1.7), we have $n \geqslant 2$. Let M be the boundary of a regular neighbourhood N of an embedding of X into $\mathbb{R}^{2n+1}$ [221, Chapter 3]. Then M is $2n$-dimensional and there is an n-connected map from M to X. We conclude from Theorem 1.7 and Theorem 8.7

$$\begin{aligned} &b_p^{(2)}(\widetilde{M}) = b_p^{(2)}(\widetilde{X}) = 0 && \text{for } p \leqslant n-1; \\ &b_p^{(2)}(\widetilde{M}) = b_{2n-p}^{(2)}(\widetilde{M}) = 0 && \text{for } p \geqslant n+1; \\ &b_n^{(2)}(\widetilde{M}) = (-1)^n \cdot \chi(M); && \\ &\alpha_p(\widetilde{M}) = \alpha_p(\widetilde{X}) = \infty^+ && \text{for } p \leqslant n; \\ &\alpha_p(\widetilde{M}) = \alpha_{2n+1-p}(\widetilde{X}) = \infty^+ && \text{for } p \geqslant n+1. \end{aligned}$$

Poincaré duality applied to (M, N), the fact that N is homotopy equivalent to X, and Theorem 1.7.2 imply

$$\chi(M) = 2 \cdot \chi(N) = 2 \cdot \chi(X) = 0.$$

Hence M is a closed Riemannian manifold satisfying

$$\begin{aligned} &b_p^{(2)}(\widetilde{M}) = 0, && p \geqslant 0; \\ &\alpha_p(\widetilde{M}) = \infty^+, && p \geqslant 0; \end{aligned}$$

(3) $\Leftrightarrow$ (4) follows from [148, Lemma 2.5, p. 31].

(4) $\Leftrightarrow$ (5) follows from [157, remark after Definition 3.11, Theorem 6.1]. □

CONJECTURE 11.4 (*Zero-in-the-spectrum Conjecture for a group*). For any finitely presented group Γ the five equivalent assertions of Lemma 11.3 are true.

REMARK 11.5. It makes sense to formulate a version of Conjecture 11.4 for arbitrary groups Γ, where one has to substitute the universal coverings $\widetilde{M}$ (respectively $\widetilde{X}$) by regular coverings of the closed manifold M (respectively the finite CW-complex X) with Γ as group of deck transformations. Of course, one then has to drop the condition that Γ is the fundamental group of M (respectively X). Also a lot of the following results can be reformulated for arbitrary groups Γ. It is likely that this more general version is true if Conjecture 11.4 holds, because the decisive condition seems to be the finiteness of X, not that $\widetilde{X}$ is simply-connected. Moreover, one may weaken the condition that X is finite to the condition that X is of finite type.

LEMMA 11.6. *Let X be a connected finite CW-complex with fundamental group Γ which is a counterexample to the zero-in-the-spectrum Conjecture* 11.4 *for the group Γ. Then*

(1) *Γ is 2-$\mathcal{N}(\Gamma)$-connected*;

(2) $\chi(X) = 0$;

(3) *If X is a closed manifold, its signature is trivial*;

(4) *If X is a closed Riemannian manifold, then $\widetilde{X}$ is not hyper-Euclidean, where hyper-Euclidean means that there is a proper distance non-increasing map from $\widetilde{X}$ to $\mathbb{R}^{\dim(X)}$ of nonzero degree. In particular, $\widetilde{X}$ and hence X, do not admit Riemannian metrics with non-positive sectional curvature*;

(5) *If X is an oriented closed manifold and $f: X \to M$ a map to an oriented closed manifold of the same dimension as X which has non-zero degree and induces an isomorphism on the fundamental groups, then M is also a counterexample to the zero-in-the-spectrum Conjecture* 11.4 *for the group Γ*;

(6) *If $X \to E \to B$ is a fibration of connected finite CW-complexes and the inclusion of X into E induces an injection on the fundamental groups, then E is a counterexample to the zero-in-the-spectrum Conjecture* 11.4 *for $\pi_1(E)$*.

PROOF. (1) The classifying map $f: \widetilde{X} \to E\Gamma$ is a Γ-equivariant 2-connected map and induces an isomorphism $H_p^{\Gamma}(\widetilde{X}; \mathcal{N}(\Gamma)) \to H_p^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma))$ for $p = 0, 1$ and an epimorphism for $p = 2$.

(2) This follows from Theorem 1.7.

(3) This is shown as in the proof of Lemma 6.5.

(4) This is proven in Gromov [109, section 8] and in [144, Proposition 7].

(5) If d is the degree of f and n the dimension of X and M, then the following diagram commutes and has isomorphisms as vertical maps by Poincaré duality

$$\begin{array}{ccc} H_p^{\Gamma}(\widetilde{X}; \mathcal{N}(\Gamma)) & \xrightarrow{\tilde{f}_*} & H_p^{\Gamma}(\widetilde{M}; \mathcal{N}(\Gamma)) \\ {\scriptstyle \cap[X]}\uparrow\cong & & {\scriptstyle d\cdot\cap[M]}\uparrow\cong \\ H_{\Gamma}^{n-p}(\widetilde{X}; \mathcal{N}(\Gamma)) & \xleftarrow{\tilde{f}^*} & H_{\Gamma}^{n-p}(\widetilde{M}; \mathcal{N}(\Gamma)) \end{array}$$

Hence the upper horizontal map is split-surjective and the claim follows.

(6) This is proven by a spectral sequence argument analogous to the proof of Theorem 10.10. □

Hence the zero-in-the-spectrum Conjecture 11.4 is true for any finitely presented group Γ which is not 2-$\mathcal{N}(\Gamma)$-connected. The next result collects some information about such groups.

DEFINITION 11.7 (*Lott* [144, *Definition* 8 *in Section* 5.1]). A finitely presented group Γ is called *big*, if Γ is non-amenable, $b_1^{(2)}(E\Gamma) = 0$ and $\alpha_2(E\Gamma) = \infty^+$. It is called *small* if it is not big.

LEMMA 11.8.

(1) *Let Γ be a group and n an integer such that $B\Gamma$ has finite $(n+1)$-skeleton. Then Γ is n-$\mathcal{N}(\Gamma)$-connected if and only if*

$$b_p^{(2)}(E\Gamma) = 0, \qquad 0 \leqslant p \leqslant n;$$
$$\alpha_p(E\Gamma) = \infty^+, \quad 0 \leqslant p \leqslant n+1;$$

(2) *A finitely presented group* Γ *is big in the sense of Lott's Definition* 11.7 *if and only if* Γ *is* 1-$\mathcal{N}(\Gamma)$*-connected. A finitely presented group* Γ *is non-amenable if and only if it is* 0-$\mathcal{N}(\Gamma)$*-acyclic;*

(3) *The fundamental group of a compact* 2*-manifold is small;*

(4) *The fundamental group of a compact connected orientable* 3*-manifold, which satisfies the assumptions of Theorem* 3.3 *that none of its prime factors is exceptional, is small;*

(5) *Let* Δ *be a normal subgroup of* Γ*. If* Δ *is* p-$\mathcal{N}(\Delta)$*-connected (respectively* $\mathcal{N}(\Delta)$*-acyclic), then* Γ *is* p-$\mathcal{N}(\Gamma)$*-connected (respectively* $\mathcal{N}(\Gamma)$*-acyclic). In particular, a finitely presented normal subgroup of a small finitely presented group is small;*

(6) *Let* Γ_0 *and* Γ_1 *be non-trivial groups. Then* $\Gamma_0 * \Gamma_1$ *is not* 1-$\mathcal{N}(\Gamma_0 * \Gamma_1)$*- connected;*

(7) *Let* Γ *be finitely presented. If* Γ *is big, then we get for the integer (see* [115])

$$q(\Gamma) = \min\{\chi(M) \mid M \text{ connected closed oriented 4-manifold with } \pi_1(M) = \Gamma\}$$

and the deficiency $\operatorname{def}(\Gamma)$

$$\operatorname{def}(\Gamma) \leqslant 1; \qquad q(\Gamma) \geqslant 0.$$

If Γ *satisfies the zero-in-the-spectrum Conjecture* 11.4*, then*

$$\operatorname{def}(\Gamma) \leqslant 0; \qquad q(\Gamma) \geqslant 1.$$

PROOF. (1) This follows from [157, remark after Definition 3.11, Theorem 6.1].

(2) This follows from (1), Theorem 1.7.9 and Theorem 8.7.6.

(3) This follows from [144, Proposition 12].

(4) This follows from [148] as carried out in [144, Proposition 13].

(5) This is proven by a spectral sequence argument analogously to the proof of Theorem 10.10.

(6) We abbreviate $\Gamma = \Gamma_0 * \Gamma_1$. Assume that Γ_0 and Γ_1 are non-trivial and Γ is 1-$\mathcal{N}(\Gamma)$-connected. A model for $B\Gamma$ is the wedge of $B\Gamma_1$ and $B\Gamma_2$. Hence one obtains a pushout of Γ-spaces

$$\begin{array}{ccc} \Gamma & \longrightarrow & \Gamma \times_{\Gamma_1} E\Gamma_1 \\ \downarrow & & \downarrow \\ \Gamma \times_{\Gamma_0} E\Gamma_0 & \longrightarrow & E\Gamma \end{array}$$

The low-dimensional part of the associated Mayer–Vietoris sequence looks like

$$\begin{aligned} \cdots &\to H_1^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma)) \to l^2(\Gamma) \\ &\to H_0^{\Gamma}(\Gamma \times_{\Gamma_0} E\Gamma_0; \mathcal{N}(\Gamma)) \oplus H_0^{\Gamma}(\Gamma \times_{\Gamma_1} E\Gamma_1; \mathcal{N}(\Gamma)) \\ &\to H_0^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma)). \end{aligned}$$

As Γ is by assumption 1-$\mathcal{N}(\Gamma)$-connected, we get an isomorphism

$$l^2(\Gamma) \xrightarrow{\cong} H_0^{\Gamma}\big(\Gamma \times_{\Gamma_0} E\Gamma_0; \mathcal{N}(\Gamma)\big) \oplus H_0^{\Gamma}\big(\Gamma \times_{\Gamma_1} E\Gamma_1; \mathcal{N}(\Gamma)\big).$$

If one applies $\dim_{\mathcal{N}(\Gamma)}$, then Theorem 10.8 implies

$$1 = b_0^{(2)}(\Gamma_0) + b_0^{(2)}(\Gamma_1).$$

Theorem 1.7.9 extends to arbitrary groups [160, Section 3]. Hence both Γ_0 and Γ_1 are of order 2 and Γ is $\mathbb{Z}/2 * \mathbb{Z}/2$. This group contains $\mathbb{Z}$ as a normal subgroup of index 2. We conclude from Example 8.5 and Theorem 8.7.5 that $\alpha_1(\Gamma) = 1$, a contradiction.

(7) Theorem 6.6 gives the first two inequalities. The improved ones in the case that Γ satisfies the zero-in-the-spectrum Conjecture are proven in [144, Sections 5.2 and 5.3]. □

REMARK 11.9. In view of Lemma 11.8.7, Lott has conjectured that $\operatorname{def}(\Gamma) \leqslant 0$ and $q(\Gamma) \geqslant 1$ holds for any finitely presented group Γ which is big [144, Conjecture 1 in 5.2 and Conjecture 2 in 5.3]. We remark that it suffices to prove this conjecture for finitely presented 2-$\mathcal{N}(\Gamma)$-connected groups. Namely, suppose that the finitely presented group Γ satisfies $b_1^{(2)}(\Gamma) = 0$ and $H_2^{\Gamma}(E\Gamma; \mathcal{N}(\Gamma)) \neq 0$. Then $\operatorname{def}(\Gamma) \leqslant 0$ and $q(\Gamma) \geqslant 1$ follow from [157, Section 6.6]. An example of a finitely presented group Γ which is big, but not 2-$\mathcal{N}(\Gamma)$-acyclic, is $(\mathbb{Z} * \mathbb{Z}) \times (\mathbb{Z} * \mathbb{Z})$.

EXAMPLE 11.10. If Γ_k is an n_k-$\mathcal{N}(\Gamma_k)$-connected group for $k = 0, 1$, then $\Gamma_0 \times \Gamma_1$ is $(n_0 + n_1)$-$\mathcal{N}(\Gamma_0 \times \Gamma_1)$-connected. This follows from the fact that the canonical $\mathbb{Z}[\Gamma_0 \times \Gamma_1]$-chain map

$$C_*(E\Gamma_0) \otimes_{\mathbb{Z}} C_*(E\Gamma_1) \to C_*\big(E(\Gamma_0 \times \Gamma_1)\big)$$

is an isomorphism. The (not finitely generated) group $\Gamma = \prod_{i=1}^{\infty} (\mathbb{Z} * \mathbb{Z})$ is $\mathcal{N}(\Gamma)$-acyclic because of Lemma 11.8.5, because it contains for each p the normal subgroup $\Gamma_n = \prod_{i=1}^{n} (\mathbb{Z} * \mathbb{Z})$ which is n-$\mathcal{N}(\Gamma_n)$-acyclic. This shows that it is crucial in the formulation of the zero-in-the-spectrum Conjecture 11.4 that in Lemma 11.3.3 to 11.3.5, the CW-complex X in question satisfies some finiteness conditions such as being finite. It may be possible that being of finite type suffices.

For residually finite Γ the results of Section 5 can be extended to the question whether zero is in the spectrum of the Laplace operator on the universal covering $\widetilde{X}$ of a connected finite CW-complex with Γ as fundamental group. Namely, the answer to this question can be read off from the low eigenvalue distributions of the Laplace operators of the various finite coverings of X given by a tower of coverings [156, Theorem 0.2, p. 456].

12. Miscellaneous

In this section we briefly mention some further aspects of L^2-invariants and give references for the reader who wants to know more about them.

The L^2-version of the index theorem was proven by Atiyah [3]. Let P be an elliptic differential operator on a closed Riemannian manifold M and let $\overline{M} \to M$ be a regular covering of M with Γ as group of deck transformations. Then we can lift the Riemannian metric and the operator P to $\overline{M}$. The operator P is Fredholm and its *index* is defined by

$$\operatorname{ind}(P) = \dim_{\mathbb{C}}\bigl(\ker(P)\bigr) - \dim_{\mathbb{C}}\bigl(\ker\bigl(P^*\bigr)\bigr).$$

Using the von Neumann trace one can define the L^2-*index* of the lifted operator $\overline{P}$ analogously

$$\operatorname{ind}_{\mathcal{N}(\Gamma)}\bigl(\overline{P}\bigr) = \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(\overline{P}\bigr)\bigr) - \dim_{\mathcal{N}(\Gamma)}\bigl(\ker\bigl(\overline{P}^*\bigr)\bigr).$$

Then the L^2-index theorem says

$$\operatorname{ind}_{\mathcal{N}(\Gamma)}\bigl(\overline{P}\bigr) = \operatorname{ind}(P).$$

If one puts elliptic boundary conditions on the operator, this result was generalized to the case where M is compact and has a boundary by Schick [228]. This generalization is the L^2-version of the index theorem in [4]. These boundary conditions are local. There are also versions of the index theorem for manifolds with boundary using global boundary conditions which apply in contrast to the local conditions, to important geometrically defined operators such as the signature operator due to Atiyah, Patodi and Singer [5–7]. This index theorem involves, as a correction term, the eta-invariant. The L^2-version of the eta-invariant is defined and studied by Cheeger and Gromov [52,53]. The L^2-version of this index theorem for manifolds with boundary and global boundary conditions is proven by Ramachandran [209] for Dirac type operators. Further references on L^2-index theory are [8,10,37,38,57,58,63–65,78,86,133,180,185,216–218,233,237,238].

Of course L^2-cohomology is not only of interest for regular coverings of closed manifolds or CW-complexes of finite type. See for instance [30,49,145,174]. In particular, the Cheeger–Goresky–MacPherson Conjecture [51] and the Zucker Conjecture [256] have created a lot of activity. They link the L^2-cohomology of the regular part with the intersection homology of an algebraic variety. References on this topic are [29,31,123,141, 192–194,198,199,204,205,222–226,255,257–259].

Connections of L^2-cohomology and discrete series of representations of Lie groups are investigated, for instance, in [8,65,102,103,230].

One can also define and investigate L^p-cohomology, as done for instance by Gromov [109, Section 8] and Pansu [201,202].

References

[1] J. Amorós et al., *Fundamental Groups of Compact Kähler Manifolds*, Mathematical Surveys and Monographs, Vol. 44, Amer. Math. Soc., Providence, RI (1996).

[2] M. Anderson, *L^2-harmonic forms and a conjecture of Singer*, Bull. Amer. Math. Soc. **13** (1985), 163–165.

[3] M. Atiyah, *Elliptic operators, discrete groups and von Neumann algebras*, Astérisque **32** (1976), 43–72.

[4] M. Atiyah and R. Bott, *The index problem for manifolds with boundary*, Differential Analysis (papers presented at the Bombay Colloquium 1994), Oxford University Press, Oxford (1964), 175–186.
[5] M. Atiyah, V.K. Patodi and I.M. Singer, *Spectral asymmetry and Riemannian geometry I*, Math. Proc. Camb. Phil. Soc. **77** (1975), 43–69.
[6] M. Atiyah, V.K. Patodi and I.M. Singer, *Spectral asymmetry and Riemannian geometry II*, Math. Proc. Camb. Phil. Soc. **78** (1975), 405–432.
[7] M. Atiyah, V.K. Patodi and I.M. Singer, *Spectral asymmetry and Riemannian geometry III*, Math. Proc. Camb. Phil. Soc. **79** (1975), 71–99.
[8] M. Atiyah and W. Schmid, *A geometric construction of the discrete series for semi-simple Lie groups*, Invent. Math. **42** (1977), 1–62.
[9] A. Avez, *Variétés riemanniennes sans points focaux*, C. R. Acad. Sc. Paris **270** (1970), 188–191.
[10] D. Barbasch and H. Moscovici, *L^2-index and the Selberg trace formula*, J. Funct. Anal. **53** (1983), 151–201.
[11] H. Bass, *The degree of polynomial growth of finitely generated nilpotent groups*, Proc. Lond. Math. Soc. **25** (1972), 603–614.
[12] H. Bass, *Euler characteristics and characters of discrete groups*, Invent. Math. **35** (1976), 155–196.
[13] P. Baum, A. Connes and R. Higson, *Classifying space for proper actions and K-theory of group C^*-algebras*, C^*-Algebras, Doran, ed., Contemp. Math., Vol. 167 (1994), 241–291.
[14] G. Baumslag, *Automorphism groups of residually finite groups*, J. London Math. Soc. **38** (1963), 117–118.
[15] G. Baumslag and P.B. Shalen, *Amalgamated products and finitely presented groups*, Comment. Math. Helv. **65** (1990), 243–254.
[16] M.E.B. Bekka and A. Valette, *Group cohomology, harmonic functions and the first L^2-Betti number*, Preprint, Potential Analysis (1994), to appear.
[17] R. Benedetti and C. Pertonio, *Lectures on Hyperbolic Geometry*, Universitext, Springer (1992).
[18] J.-M. Bismut, *From Quillen metrics to Reidemeister metrics: some aspects of the Ray–Singer analytic torsion*, Topological Methods in Modern Mathematics, A Symposium in Honor of John Milnor's Sixtieth Birthday, Publish or Perish, Houston (1993), 273–324.
[19] J.-M. Bismut and D.S. Freed, *The analysis of elliptic families: Metrics and connections on determinant bundle*, Comm. Math. Phys. **106** (1986), 159–176.
[20] J.-M. Bismut and D.S. Freed, *The analysis of elliptic families: Dirac operators, eta invariants and the holonomy theorem of Witten*, Comm. Math. Phys. **107** (1986), 103–163.
[21] J.M. Bismut, H. Gillet and C. Soulé, *Analytic torsion and holomorphic determinant bundles I: Bott Chern forms and analytic torsion*, Comm. Math. Phys. **115** (1988), 49–78.
[22] J.M. Bismut, H. Gillet and C. Soulé, *Analytic torsion and holomorphic determinant bundles II: Direct images and Bott Chern forms*, Comm. Math. Phys. **115** (1988), 79–126.
[23] J.M. Bismut, H. Gillet and C. Soulé, *Analytic torsion and holomorphic determinant bundles III: Quillen metrics on holomorphic determinants*, Comm. Math. Phys. **115** (1988), 301–351.
[24] J.M. Bismut and J. Lott, *Flat vector bundles, direct images and higher real analytic torsion*, J. Amer. Math. Soc. **8** (1995), 291–363.
[25] J.M. Bismut and W. Zhang, *An extension of a theorem by Cheeger and Müller*, Astérisque **205** (1992).
[26] J.M. Bismut and W. Zhang, *Milnor and Ray–Singer metrics on the equivariant determinant of a flat vector bundle*, GAFA **4** (1994), 136–212.
[27] M. Boileau and S. Druck, *Compact foliations of open manifolds with amenable leaves*, Preprint, Toulouse (1991).
[28] M. Boileau, S. Druck and E. Vogt, *Vanishing of Gromov volume and circle foliations of Epstein length* 1 *on open* 3*-manifolds*, Preprint, Toulouse (1995).
[29] A. Borel, *The L^2-cohomology of negatively curved Riemannian symmetric spaces*, Ann. Scient. Ec. Norm. Sup. 4ème série **7** (1974), 235–272.
[30] A. Borel and W. Casselmann, *The L^2-cohomology of locally symmetric manifolds of finite volume*, Duke Math. J. **50** (1983), 625–647.
[31] A. Borel and W. Casselmann, *Cohomology d'intersection et L^2-cohomology de variétés arithmetiques de rang rationnel* 2, C. R. Acad. Sci. Paris Sér. Math. I **301** (1985), 369–373.
[32] M. Bravermann, *Witten deformation and analytic torsion and the spectral sequence of a fibration*, GAFA **6** (1996), 28–50.

[33] M.G. Brin and C.C. Squier, *Groups of piecewise linear homeomorphisms of the real line*, Invent. Math. **79** (1985), 495–498.

[34] R. Brooks, *The fundamental group and the spectrum of the Laplacian*, Comment. Math. Helv. **56** (1981), 581–598.

[35] K.S. Brown, *Cohomology of Groups*, Graduate Texts in Math., Vol. 87, Springer, Berlin (1982).

[36] K.S. Brown and R. Geoghegan, *An infinite-dimensional torsion-free FP_∞ group*, Invent. Math. **77** (1984), 367–381.

[37] J. Brüning, *L^2-index theorems for certain complete manifolds*, J. Differential Geom. **32** (1990), 491–532.

[38] J. Brüning, *L^2-index theorems for complete manifolds of rank-one type*, Duke Math. J. **66** (1992), 491–532.

[39] J. Brüning and M. Lesch, *Hilbert complexes*, J. Funct. Anal. **108** (1992), 88–132.

[40] U. Bunke, *Equivariant higher analytic torsion und equivariant Euler characteristic*, Preprint, Göttingen (1997).

[41] D. Burghelea, L. Friedlander and T. Kappeler, *Torsion for manifolds with boundary and glueing formulas*, Preprint (1996).

[42] D. Burghelea, L. Friedlander and T. Kappeler, *Witten deformation and analytic torsion*, Preprint, to appear in a volume for Krein (1996).

[43] D. Burghelea, L. Friedlander, T. Kappeler and P. McDonald, *Analytic and Reidemeister torsion for representations in finite type Hilbert modules*, Geom. Funct. Anal. **6** (1996), 751–859.

[44] D. Burghelea, L. Friedlander, T. Kappeler and P. McDonald, *Asymptotic expansion of the Witten deformation of analytic torsion*, J. Funct. Anal. **137** (1996), 320–363.

[45] J.W. Cannon, W.J. Floyd and W.R. Parry, *Notes on Richard Thompson's groups F and T*, Research Report GCG63, The Geometry Center, Univ. Minn. (1994).

[46] S. Cappell and J.L. Shaneson, *Non-linear similarity*, Ann. of Math. **113** (1981), 315–355.

[47] A.L. Carey and V. Mathai, *L^2-torsion invariants*, J. Funct. Anal. **110** (1992), 377–409.

[48] A.L. Carey, M. Farber and V. Mathai, *Determinant lines, von Neumann algebras and L^2-torsion*, J. Reine Angew. Math. **484** (1997), 153–181.

[49] W. Casselmann, *An introduction to the L^2-cohomology of arithmetic quotients of bounded symmetric domains*, Complex Analytic Singularities, Advanced Studies in Pure Mathematics, Vol. 8, North Holland, Amsterdam (1987), 69–93.

[50] J. Cheeger, *Analytic torsion and the heat equation*, Annals of Math. **109** (1979), 259–322.

[51] J. Cheeger, M. Goresky and R. McPherson, *$L_{(2)}$-cohomology and intersection homology for singular algebraic varieties*, Seminar on Differential Geometry, S. Yau, ed., Annals of Math. Studies, Vol. 102, Princeton (1982), 303–346.

[52] J. Cheeger and M. Gromov, *Bounds on the von Neumann dimension of L^2-cohomology and the Gauss–Bonnett theorem for open manifolds*, J. Differential Geom. **21** (1985), 1–34.

[53] J. Cheeger and M. Gromov, *On the characteristic numbers of complete manifolds of bounded curvature and finite volume*, Rauch Mem., Vol., I. Chavel and H. Farkas, eds, Springer, Berlin (1985), 115–154.

[54] J. Cheeger and M. Gromov, *L^2-cohomology and group cohomology*, Topology **25** (1986), 189–215.

[55] P. Chernoff, *Essential self-adjointness of powers of generators of hyperbolic equations*, J. Funct. Anal. **12** (1973), 401–414.

[56] B. Clair, *Residual amenability and the approximation of L^2-invariants*, Ph.D. thesis (1997).

[57] J. Cohen, *Von Neumann dimension and the homology of covering spaces*, Quart. J. Math. Oxford **30** (1979), 133–142.

[58] J. Cohen, *L^2-cohomology in topology, algebra and analysis*, Rend. Circ. Mat. Palermo (2) suppl. **18** (1988), 31–36.

[59] M.M. Cohen, *A Course in Simple Homotopy Theory*, Graduate Texts in Math., Vol. 10, Springer, Berlin (1973).

[60] M.M. Cohen, *Combinatorial Group Theory: A Topological Approach*, LMS Student Texts, Vol. 14 (1989).

[61] P.M. Cohn, *Free Rings and their Relations* (2nd edition), LMS Monographs, Vol. 19, Academic Press (1985).

[62] A. Connes, *Non-commutative differential topology*, Publ. Math. IHES **62** (1985), 257–360.

[63] A. Connes, *Non-commutative Geometry*, Academic Press (1994).

[64] A. Connes and H. Moscovici, *The L^2-index theorem for homogeneous spaces of Lie groups*, Ann. of Math. **115** (1982), 291–330.

[65] A. Connes and H. Moscovici, *The L^2-index theory on homogeneous spaces and discrete series representations*, Operator Algebras and Applications, Proc. Symp. Pure Math. 38, Vol. 1, Amer. Math. Soc., Providence, RI (1982), 419–433.

[66] P. Connor and F. Raymond, *Actions of compact Lie groups on aspherical manifolds*, Topology of Manifolds, Cantrell and Edwards, Markham, Chicago (1970), 227–264.

[67] X. Dai and R.B. Melrose, *Adiabatic limit of the analytic torsion*, Preprint (1995).

[68] J. Davis and W. Lück, *Spaces over a category and assembly maps in isomorphism conjectures in K- and L-theory*, K-theory **15** (1998), 201–252.

[69] A. Deitmar, *A determinant formula for the generalized Selberg zeta function*, Quart. J. Math. Oxford **2** (1996), 435–453.

[70] A. Deitmar, *Regularized and L^2-determinants*, Proc. London Math. Soc. **76** (1998), 150–174.

[71] A. Deitmar, *Equivariant torsion of locally symmetric spaces*, Pacific J. Math. **182** (1998), 205–227.

[72] J. Dodziuk, *DeRham–Hodge theory for L^2-cohomology of infinite coverings*, Topology **16** (1977), 157–165.

[73] J. Dodziuk, *L^2-harmonic forms on rotationally symmetric Riemannian manifolds*, Proc. Amer. Math. Soc. **77** (1979), 395–400.

[74] J. Dodziuk and V. Mathai, *Approximating L^2-invariants of amenable covering spaces: A combinatorial approach*, J. Funct. Anal. **154** (1998), 359–378.

[75] H. Donnelly, *The differential form spectrum of hyperbolic space*, Manuscripta Math. **33** (1981), 365–385.

[76] H. Donnelly, *On L^2-Betti numbers for abelian groups*, Can. Math. Bull. **24** (1981), 91–95.

[77] H. Donnelly, *On the spectrum of towers*, Proc. Amer. Math. Soc. **87** (1983), 322–329.

[78] H. Donnelly and C. Fefferman, *L^2-cohomology and index theorem for the Bergman metric*, Ann. of Math. **118** (1983), 593–618.

[79] H. Donnelly and F. Xavier, *On the differential form spectrum of negatively curved Riemannian manifolds*, Amer. J. Math. **106** (1984), 169–185.

[80] B. Eckmann, *Amenable groups and Euler characteristic*, Comment. Math. Helv. **67** (1992), 383–393.

[81] B. Eckmann, *Manifolds of even dimension with amenable fundamental groups*, Comment. Math. Helv. **69** (1994), 501–511.

[82] B. Eckmann, *Finitely dominated spaces and l^2-Euler characteristic*, Preprint, Zürich (1995).

[83] E.G. Effros, *Why the circle is connected: An introduction to quantized topology*, Math. Intelligencer **11** (1989), 27–34.

[84] A. Efremov, *Cell-decompositions and the Novikov–Shubin invariants*, Russ. Math. Surv. **46** (1991), 219–220.

[85] A. Efremov, *Combinatorial and analytic Novikov–Shubin invariants*, Preprint (1991).

[86] J. Eichhorn, *L^2-cohomology*, Geometric and Algebraic Topology, Banach Center Publ., Vol. 18, 215–239.

[87] D.B.A. Epstein, *Finite presentations of groups and 3-manifolds*, Quart. J. Math. Oxford II Ser. **12** (1961), 205–212.

[88] M.S. Farber, *Singularities of the analytic torsion*, J. Differential Geom. **41** (1995), 528–572.

[89] M.S. Farber, *Abelian categories, Novikov–Shubin invariants, and Morse inequalities*, C. R. Acad. Sci. Paris **321** Sèrie I (1995), 1593–1598.

[90] M.S. Farber, *Homological algebra of Novikov–Shubin invariants and Morse inequalities*, GAFA **6** (1996), 628–665.

[91] M.S. Farber, *Von Neumann categories and extended L^2-cohomology*, Preprint, IHES (1996).

[92] M. Farber, *Geometry of growth: Approximation theorems for L^2-invariants*, Math. Ann. **311** (1998), 335–376.

[93] F.T. Farrell and L.E. Jones, *Isomorphism conjectures in algebraic K-theory*, J. Amer. Math. Soc. **6** (1993), 249–298.

[94] W. Franz, *Über die Torsion einer Überdeckung*, J. Reine Angew. Math. **173** (1935), 245–254.

[95] D. Fried, *Analytic torsion and closed geodesics on hyperbolic manifolds*, Invent. Math. **84** (1986), 523–540.

[96] D. Fried, *Torsion and closed geodesics on complex hyperbolic manifolds*, Invent. Math. **91** (1988), 31–51.

[97] B. Fuglede and R.V. Kadison, *Determinant theory in finite factors*, Ann. of Math. **55** (1952), 520–530.

[98] D. Gabai, *Convergence groups are Fuchsian groups*, Bull. Amer. Math. Soc. **25** (1991), 395–402.

[99] S. Gallot, D. Hulin and J. Lafontaine, *Riemannian Geometry*, Universitext, Springer, Berlin (1987).

[100] P. Gilkey, *Invariance Theory, the Heat Equation and the Atiyah–Singer Index Theorem*, Cambridge University Press (1994).

[101] H. Gillet and C. Soulé, *Analytic torsion and the arithmetic todd genus*, Topology **30** (1991), 21–54.

[102] D.L. deGorge and N.R. Wallach, *Limit multiplicities in $L^2(\Gamma\backslash G)$ I*, Ann. of Math. **107** (1978), 133–150.

[103] D.L. deGorge and N.R. Wallach, *Limit multiplicities in $L^2(\Gamma\backslash G)$ II*, Ann. of Math. **109** (1979), 477–495.

[104] D. Gottlieb, *A certain subgroup of the fundamental group*, Amer. J. Math. **87** (1965), 840–856.

[105] M. Gromov, *Groups of polynomial growth and expanding maps*, Publ. Math. IHES **53** (1981), 53–73.

[106] M. Gromov, *Volume and bounded cohomology*, Publ. Math. IHES **56** (1982), 5–100.

[107] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, Math. Sci. Res. Ins. Publ., Vol. 8, Springer, Berlin (1987), 75–263.

[108] M. Gromov, *Kähler hyperbolicity and L_2-Hodge theory*, J. Differential Geom. **33** (1991), 263–292.

[109] M. Gromov, *Asymptotic invariants of infinite groups*, Geometric Groups Theory, Vol. 2, Proc. of the Symp. in Sussex 1991, G.A. Niblo and M.A. Roller, eds, Lecture Notes Series, Vol. 182, Cambridge University Press (1993).

[110] M. Gromov and M.A. Shubin, *Von Neumann spectra near zero*, GAFA **1** (1991), 375–404.

[111] M. Gromov and M.A. Shubin, *Near-cohomology of Hilbert spaces and topology of non-simply connected manifolds*, Astérisque **210** (1992), 283–294.

[112] M. Gromov and M.A. Shubin, *Erratum to Von Neumann spectra near zero*, GAFA **5** (1995), 729.

[113] K.W. Grünberg, *Residual properties of infinite soluble groups*, Proc. London Math. Soc. **7** (1957), 29–62.

[114] I. Hambleton and E.K. Pedersen, *Non-linear similarity revisited*, Prospects in Topology, Proceedings of a Conference in Honour of William Browder, Princeton, March 1994, Annals of Math. Studies, Vol. 138, Princeton University Press (1995), 157–174.

[115] J.-C. Hausmann and S. Weinberger, *Caractéristiques d'Euler et groupes fondamentaux des variét'es de dimension* 4, Comment. Math. Helv. **60** (1985), 139–144.

[116] J. Hempel, *3-Manifolds*, Annals of Math. Studies, Vol. 86, Princeton University Press (1976).

[117] J. Hempel, *Residual finiteness for 3-manifolds*, Combinatorial Group Theory and Topology, S.M. Gersten and J.R. Stallings, eds, Annals of Math. Studies, Vol. 111, Princeton University Press, Princeton (1987), 379–396.

[118] E. Hess and T. Schick, *Nonvanishing of L^2-torsion of hyperbolic manifolds*, Preprintreihe SFB 478, Heft 18, Manuscripta Math. (1998), to appear.

[119] G. Higman, *A finitely generated infinite simple group*, J. London Math. Soc. **26** (1951), 61–64.

[120] J.A. Hillman, *Elementary amenable groups and 4-manifolds with Euler characteristic* 0, J. Australian Math. Soc. **50** (1991), 160–170.

[121] J.A. Hillman, *L^2-homology and asphericity*, Israel J. Math. **99** (1997), 271–283.

[122] C. Hog, M. Lustig and W. Metzler, *Presentation classes, 3-manifolds and free products*, Proc. Conference "Geometry and Topology" 1983–1984, University of Maryland, J. Alexander and J. Harer, eds, Lecture Notes in Math., Vol. 1167 (1985), 154–167.

[123] W.C. Hsiang and C. Pati, *L^2-cohomology of normal algebraic surfaces*, Invent. Math. **81** (1985), 395–412.

[124] W.C. Hsiang and W. Pardon, *When are topologically equivalent orthogonal representations linearly equivalent*, Invent. Math. **38** (1982), 275–317.

[125] W. Jaco and P.B. Shalen, *Seifert fibred spaces in 3-manifolds*, Mem. Amer. Math. Soc. **220** (1979).

[126] N. Ivanov, *Foundations of the theory of bounded cohomology*, J. Soviet Math. **37** (1987), 1090–1114.

[127] K. Johannson, *Homotopy Equivalences of 3-Manifolds with Boundary*, Lecture Notes in Math., Vol. 761, Springer, Berlin (1979).

[128] F.E.A. Johnson and D. Kotschick, *On the signature and Euler characteristic of certain four-manifolds*, Math. Proc. Camb. Phil. Soc. **114** (1993), 431–437.

[129] J. Jost and L.X. Yuanlong, *Vanishing theorems for L^2-cohomology groups*, Preprint (1997).

[130] J. Jost and K. Zuo, *Vanishing theorems for L^2-cohomology on infinite coverings of compact Kähler manifolds and applications in algebraic geometry*, Preprint (1997).

[131] R.V. Kadison and J.R. Ringrose, *Fundamentals of the Theory of Operator Algebras, Vol. II: Advanced Theory*, Pure and Applied Mathematics, Academic Press (1986).

[132] D. Kazhdan, *On arithmetic varieties*, Lie Groups and their Representations, Halsted (1975), 151–216.

[133] N. Keswani, *Relative Eta-invariants and C^*-algebra K-theory*, Preprint, Preprintreihe SFB 478, Heft 26 (1998).
[134] K. Kodaira, *On Kähler varieties of restricted type (an intrinsic characterization of algebraic varieties)*, Ann. of Math. **60** (1954), 28–48.
[135] K. Köhler, *Holomorphic torsion on Hermitian symmetric spaces*, J. Reine Angew. Math. **460** (1995), 93–116.
[136] K. Köhler, *Equivariant Reidemeister torsion on symmetric spaces*, Math. Ann. **307** (1997), 57–69.
[137] M. Kreck, W. Lück and P. Teichner, *Counterexamples to the Kneser conjecture in dimension four*, Comment. Math. Helv. **70** (1995), 423–433.
[138] T.Y. Lam, *A First Course in Noncommutative Rings*, Graduate Texts in Math., Vol. 131, Springer, Berlin (1991).
[139] E.C. Lance, *On nuclear C^*-algebras*, J. Funct. Anal. **12** (1973), 157–176.
[140] P. Linnell, *Division rings and group von Neumann algebras*, Forum Math. **5** (1993), 561–576.
[141] E. Looijenga, *L^2-cohomology of locally symmetric varieties*, Comp. Math. **67** (1988), 3–20.
[142] J. Lott, *Heat kernels on covering spaces and topological invariants*, J. Differential Geom. **35** (1992), 471–510.
[143] J. Lott, *Equivariant analytic torsion for compact Lie group actions*, J. Funct. Anal. **125** (1994), 438–451.
[144] J. Lott, *The zero-in-the-spectrum question*, L'Enseignement Mathématique **42** (1996), 341–376.
[145] J. Lott, *L^2-cohomology of geometrically infinite hyperbolic 3-manifolds*, Geom. Funct. Anal. **7** (1997), 81–119.
[146] J. Lott, *Delocalized L^2-invariants*, Preprint, J. Funct. Anal. (1996), to appear.
[147] J. Lott, *Deficiencies of lattice subgroups of Lie groups*, Preprint, Bull. London Math. Soc. (1997), to appear.
[148] J. Lott and W. Lück, *L^2-topological invariants of 3-manifolds*, Invent. Math. **120** (1995), 15–60.
[149] J. Lott and M. Rothenberg, *Analytic torsion for group actions*, J. Differential Geom. **34** (1991), 431–481.
[150] W. Lück, *The transfer maps induced in the algebraic K_0- and K_1-groups by a fibration I*, Math. Scand. **59** (1986), 93–121.
[151] W. Lück, *The transfer maps induced in the algebraic K_0- and K_1-groups by a fibration II*, J. Pure Appl. Algebra **45** (1987), 143–169.
[152] W. Lück, *Transformation Groups and Algebraic K-Theory*, Lecture Notes Math., Vol. 1408 (1989).
[153] W. Lück, *Analytic and topological torsion for manifolds with boundary and symmetries*, J. Differential Geom. **37** (1993), 263–322.
[154] W. Lück, *L^2-torsion and 3-manifolds*, Conference Proceedings and Lecture Notes in Geometry and Topology, Vol. III "Low-dimensional topology", Knoxville 1992, K. Johannson, ed., International Press (1994), 75–107.
[155] W. Lück, *L^2-Betti numbers of mapping tori and groups*, Topology **33** (1994), 203–214.
[156] W. Lück, *Approximating L^2-invariants by their finite-dimensional analogues*, GAFA **4** (1994), 455–481.
[157] W. Lück, *Hilbert modules and modules over finite von Neumann algebras and applications to L^2-invariants*, Math. Ann. **309** (1997), 247–285.
[158] W. Lück, *L^2-Invarianten von Mannigfaltigkeiten und Gruppen*, Jahresber. Deutsch. Math.-Verein. **99** (1997), 101–109.
[159] W. Lück, *Dimension theory of arbitrary modules over finite von Neumann algebras and L^2-Betti numbers I: Foundations*, J. Reine Angew. Math. **495** (1998), 135–162.
[160] W. Lück, *Dimension theory of arbitrary modules over finite von Neumann algebras and L^2-Betti numbers II: Applications to Grothendieck groups*, J. Reine Angew. Math. **496** (1998), 213–236.
[161] W. Lück, *L^2-Invariants and K-Theory*, in preparation, Springer.
[162] W. Lück, H. Reich and T. Schick, *Novikov–Shubin invariants for arbitrary group actions and their positivity*, Preprint, SFB-Reihe, Heft 23, Proceedings of the Topology Conference, Tel Aviv, May 1998 (1998), to appear.
[163] W. Lück and M. Rothenberg, *Reidemeister torsion and the K-theory of von Neumann algebras*, K-theory **5** (1991), 213–264.
[164] W. Lück and T. Schick, *L^2-torsion of hyperbolic manifolds of finite volume*, Preprintreihe SFB 478 – Geometrische Strukturen in der Mathematik, Münster, Heft 3, GAFA (1998), to appear.
[165] W. Lück, T. Schick and T. Thielmann, *Torsion and fibrations*, J. Reine Angew. Math. **498** (1998), 1–33.

[166] R.C. Lyndon and P.E. Schupp, *Combinatorial Group Theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 89, Springer, Berlin (1977).

[167] I. Madsen and M. Rothenberg, *On the classification of G-spheres I: Equivariant transversality*, Acta Math. **160** (1988), 65–104.

[168] A. Mal'cev, *On isomorphic matrix representations of finite groups*, Mat. Sb. **8** (1940), 405–422.

[169] W. Magnus, *Residually finite groups*, Bull. Amer. Math. Soc. **75** (1969), 305–315.

[170] V. Mathai, *L^2-analytic torsion*, J. Funct. Anal. **107** (1992), 369–386.

[171] V. Mathai, *L^2-invariants of Covering Spaces*, Geometric Analysis and Lie Theory in Mathematics and Physics, A.L. Carey, ed., Aust. Math. Soc. Lect., Vol. 11, Cambridge University Press (1998).

[172] V. Mathai and M.A. Shubin, *Twisted L^2 invariants of non-simply connected manifolds and asymptotic L^2 Morse inequalities*, Preprint (1996).

[173] V. Mathai and S.G. Yates, *Discrete Morse theory and extended L^2-homology*, Preprint (1998).

[174] R. Mazzeo and R.S. Phillips, *Hodge theory on hyperbolic manifolds*, Duke Math. J. **60** (1990), 509–559.

[175] J. Milnor, *Whitehead torsion*, Bull. Amer. Math. Soc. **72** (1966), 358–426.

[176] E.E. Moise, *Affine structures in 3-manifolds V: The triangulation theorem and the Hauptvermutung*, Ann. of Math. **55** (1952), 96–114.

[177] B.G. Moishezon, *On n-dimensional compact varieties with n algebraically independent meromorphic functions*, Amer. Math. Soc. Translations **63** (1967), 51–177.

[178] J.A. Moody, *Brauer induction for G_0 of certain infinite groups*, J. Algebra **122** (1989), 1–14.

[179] J.W. Morgan, *On Thurston's Uniformization Theorem for Three-Dimensional Manifolds*, The Smith Conjecture, J.W. Morgan and H. Bass, eds, Academic Press (1984).

[180] H. Moscovici, *L^2-index of elliptic operators in locally symmetric spaces of finite volume*, Operator Algebras and K-Theory, Contemp. Math., Vol. 10, Amer. Math. Soc., Providence, RI (1982), 129–138.

[181] H. Moscovici and R.J. Stanton, *R-torsion and zeta functions for locally symmetric manifolds*, Invent. Math. **105** (1991), 185–216.

[182] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Annals of Math. Studies, Vol. 78, Princeton University Press, Princeton (1973).

[183] A.W. Mostowski, *On the decidability of some problems in special classes of groups*, Fund. Math. **59** (1966), 123–135.

[184] W. Müller, *Analytic torsion and R-torsion of Riemannian manifolds*, Adv. in Math. **28** (1978), 233–305.

[185] W. Müller, *L^2-index theory, Eta invariants and values of L-functions*, Geometric and Topological Invariants for Elliptic Operators, Brunswick 1988, Contemp. Math., Vol. 105 (1990), 145–189.

[186] W. Müller, *Analytic torsion and R-torsion for unimodular representations*, J. Amer. Math. Soc. **6** (1993), 721–753.

[187] W. Müller, *Spectral theory and geometry*, Proc. of the First European Congress of Mathematics, July 6–10, 1992, Paris, Vol. I, Birkhäuser (1994), 153–185.

[188] F.J. Murray and J. von Neumann, *On rings of operators*, Ann. of Math. **37** (1936), 116–229.

[189] F.J. Murray and J. von Neumann, *On rings of operators II*, Trans. Amer. Math. Soc. **41** (1937), 208–248.

[190] F.J. Murray and J. von Neumann, *On rings of operators IV*, Ann. of Math. **44** (1943), 716–808.

[191] J.R. Munkres, *Obstructions to smoothing piecewise-differentiable homeomorphisms*, Ann. of Math. **72** (1960), 521–554.

[192] M. Nagase, *L_2-cohomology and intersection homology of stratified spaces*, Duke Math. J. **50** (1983), 329–368.

[193] M. Nagase, *Sheaf theoretic L^2-cohomology*, Advanced Studies in Pure Math., Vol. 8, North-Holland (1986), 273–279.

[194] M. Nagase, *Remarks on the L^2-cohomology of singular algebraic surfaces*, J. Math. Soc. Japan **41** (1989), 97–116.

[195] H. Neumann, *Varieties of Groups*, Ergebnisse der Mathematik und ihrer Grenzgebiete 37, Springer, Berlin (1967).

[196] S. Novikov and M. Shubin, *Morse inequalities and von Neumann invariants of non-simply connected manifolds*, Sov. Math. Dokl. **34** (1987), 79–82.

[197] S. Novikov and M. Shubin, *Morse inequalities and von Neumann II_1-factors*, Dokl. Akad. Nauk **34** (1) (1986), 289–292; Soviet. Math. Dokl. **34** (1) (1987), 79–82.

[198] T. Ohsawa, *On the L^2-cohomology groups of complex spaces*, Math. Z. **209** (1992), 519–530.

[199] T. Ohsawa, *On the L^2-cohomology groups of isolated singularities*, Advanced Studies in Mathematics, Vol. 22, Progress in Differential Geometry (1993), 247–263.
[200] A.J. Olšhanskii, *An infinite simple group without torsion*, Math. USSR. Izv. **15** (1980), 531–588.
[201] P. Pansu, *Cohomologie L^p des variétés à courbure négative, cas du degré un*, P.D.E. and Geometry 1988, Rend. Sem. Mat. Torino, Fasc. spez. (1989), 95–120.
[202] P. Pansu, *Cohomologie L^p: invariance sous quasiisométrie*, Preprint (1995).
[203] P. Pansu, *Introduction to L^2-Betti numbers*, Riemannian Geometry, Fields Institute Monographs, Vol. 4, M. Lovric, M. Min Oo and McK. Wang, eds, Amer. Math. Soc., Providence, RI (1996), 53–86.
[204] W. Pardon, *L^2-$\bar{\partial}$-cohomology of an algebraic surface*, Topology **28** (1989), 171–195.
[205] W. Pardon and M. Stern, *L^2-$\bar{\partial}$-cohomology of complex projective varieties*, Topology **28** (1991), 171–195.
[206] A.L. Paterson, *Amenability*, Math. Surveys and Monographs, Vol. 29, Amer. Math. Soc., Providence, RI (1988).
[207] G. Pedersen, *C^*-Algebras and their Automorphism Groups*, Academic Press (1979).
[208] D. Quillen, *Determinants of Cauchy–Riemann operators over a Riemannian surface*, Funct. Anal. Appl. **19** (1986), 31–35.
[209] M. Ramachandran, *Von Neumann index theorems for manifolds with boundary*, J. Differential Geom. **38** (1993), 315–349.
[210] J.G. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Graduate Texts in Mathematics, Vol. 149, Springer, Berlin (1994).
[211] D.B. Ray and I. Singer, *R-torsion and the Laplacian on Riemannian manifolds*, Adv. in Math. **7** (1971), 145–210.
[212] D.B. Ray and I. Singer, *Analytic torsion for complex manifolds*, Ann. of Math. **98** (1973), 154–177.
[213] H. Reich, *Group von Neumann algebras and related algebras*, Ph.D. thesis, Göttingen (1999).
[214] K. Reidemeister, *Homotopieringe und Linsenräume*, Hamburger Abhandlungen **11** (1938).
[215] G. deRham, *Reidemeister's torsion invariant and rotations of S^n*, Differential Analysis (published for the Tata Institute of Fundamental Research, Bombay), Oxford University Press, London (1964), 27–36; MR 32♯8355.
[216] J. Roe, *An index theorem on open manifolds I*, J. Differential Geom. **27** (1988), 87–113.
[217] J. Roe, *An index theorem on open manifolds II*, J. Differential Geom. **27** (1988), 115–136.
[218] J. Roe, *On the quasi-isometry invariance of L^2-Betti numbers*, Duke Math. J. **59** (1989), 765–783.
[219] S. Rosset, *A vanishing theorem for Euler characteristic*, Math. Z. **185** (1984), 211–215.
[220] M. Rothenberg, *Torsion invariants and finite transformation groups*, Algebraic and Geometric Topology, Part 1, Proc. Symp. Pure Math., Vol. 33, Amer. Math. Soc., Providence, RI (1978), 267–311.
[221] C.P. Rourke and B.J. Sanderson, *Introduction to Piecewise-Linear Topology*, Springer, Berlin (1982).
[222] L. Saper, *L_2-cohomology and intersection homology of certain algebraic varieties with isolated singularities*, Invent. Math. **82** (1985), 207–255.
[223] L. Saper, *L_2-cohomology of algebraic varieties*, Proc. ICM, Kyoto, Japan 1990, Part I, Math Soc. Japan, Springer, Berlin (1991), 735–746.
[224] L. Saper, *L_2-cohomology of Kähler varieties with isolated singularities*, J. Differential Geom. **36** (1992), 89–161.
[225] L. Saper and M. Stern, *L_2-cohomology of arithmetic varieties*, Ann. of Math. **132** (1990), 1–69.
[226] L. Saper and S. Zucker, *An introduction to L_2-cohomology*, Several Complex Variables and Complex Geometry, Santa Cruz, July 10–30, 1989, E. Bedford, J.P. D'Angelo, R.E. Greene and S.G. Krantz, eds, Proc. of Symp. in Pure Math., Vol. 52 Part 2 (1991), 519–534.
[227] P. Sarnak and X. Xue, *Bounds for the multiplicity of automorphic representations*, Duke Math J. **64** (1991), 207–227.
[228] T. Schick, *Analysis on ∂-manifolds of bounded geometry, Hodge–deRham isomorphism and L^2-index theorem*, Dissertation, Mainz (1996).
[229] T. Schick, *L^2-determinant class and approximation of L^2-Betti numbers*, Preprint, SFB 418 Münster (1998).
[230] W. Schmid, *L^2-cohomology and discrete series*, Ann. of Math. **103** (1976), 375–394.
[231] P. Scott, *The geometry of 3-manifolds*, Bull. London Math. Soc. **15** (1983), 401–487.
[232] R.T. Seeley, *Complex powers of elliptic operators*, Proc. Symp. Pure and Appl. Math., Vol. 10, Amer. Math. Soc., Providence, RI (1967), 288–307.

[233] M.A. Shubin, *A Lefschetz fixed point formula in reduced L^2-cohomologies of manifolds with cylinders*, Analysis on Manifolds with Singularities, Breitenbrunn 1990, B.W. Schulze, ed., Teubner Texts Math., Vol. 133, Teubner (1992), 290–297.

[234] M.A. Shubin, *Semiclassical asymptotics on covering manifolds and Morse inequalities (survey)*, GAFA **6** (1996), 370–409.

[235] M.A. Shubin, *DeRham theorem for extended L^2-cohomology*, Preprint (1996).

[236] T. Soma, *The Gromov invariant of links*, Invent. Math. **64** (1981), 445–454.

[237] M. Stern, *L^2-index theorems on locally symmetric spaces*, Invent. Math. **96** (1989), 231–282.

[238] M. Stern, *Index theory for certain complete Kähler manifolds*, J. Differential Geom. **37** (1993), 467–503.

[239] T. Thielmann, *Die Leray–Serre Spektralsequenz für getwistete Koeffizienten und L^2-Kohomologie*, Staatsexamensarbeit, Mainz (1993).

[240] W. Thurston, *The Geometry and Topology of Three-Manifolds*, Princeton University Lecture Notes (1978–1979).

[241] N. Varapoulos, *Random walks and Brownian motion on manifolds*, Symposia Mathematica **29** (1988), 97.

[242] S.M. Vishik, *Analytic torsion of the boundary value problems*, Dokl. Akad. Nauk SSSR **295** (1987), 1293–1298 (Soviet Math. Dokl. **36** (1988), 174–179).

[243] S.M. Vishik, *Analytic torsion for finite-dimensional local systems*, Dokl. Akad. Nauk SSSR **300** (1988), 1295–1299 (Soviet Math. Dokl. **37** (1988), 803–807).

[244] S.M. Vishik, *Generalized Ray–Singer conjecture I: A manifold with smooth boundary*, Comm. Math. Phys. **167** (1995), 1–102.

[245] J. Wagoner, *Diffeomorphisms, K_2 and analytic torsion*, Algebraic and Geometric Topology, Stanford, August 2–21, 1976, Proc. Symp. in Pure Math. of the Amer. Math. Soc., Vol. 32 Part 1 (1978), 23–34.

[246] C.T.C. Wall, *Rational Euler characteristics*, Proc. Camb. Phil. Soc. **51** (1961), 181–183.

[247] C.T.C. Wall, *Surgery on Compact Manifolds*, Academic Press (1970).

[248] B.A.F. Wehrfritz, *Infinite Linear Groups*, Ergebnisse der Mathematik und ihrer Grenzgebiete 76, Springer, Berlin (1973).

[249] R.O. Wells, *Differential Analysis on Complex Manifolds*, Graduate Texts in Mathematics, Vol. 65, Springer, Berlin (1980).

[250] H.E. Winkelkempner, *Un teorema sobre variedades de dimensión* 4, Acta Mexicana Ci. Tecn. **2** (1968), 88–89.

[251] X. Xue, *On the first Betti numbers of hyperbolic spaces*, Duke Math. J. **64** (1991), 85–110.

[252] X. Xue, *On the Betti numbers of hyperbolic manifolds*, GAFA **2** (1993), 126–136.

[253] K. Yano, *Gromov invariant and S^1-action*, J. Fac. Sc. U. Tokyo, Sec. 1A Math. **29** (1982), 493–501.

[254] S.K. Yeung, *Betti numbers on a tower of coverings*, Duke Math. J. **73** (1994), 201–226.

[255] S. Zucker, *$L_{(2)}$-cohomology of warped products and arithmetic groups*, Invent. Math. **70** (1982), 169–218.

[256] S. Zucker, *L^2-cohomology and intersection homology of locally symmetric varieties*, Proc. Symp. Pure Math., Vol. 40, Part 2, Amer. Math. Soc., Providence, RI (1983), 675–680.

[257] S. Zucker, *L^2-cohomology and intersection homology of locally symmetric varieties II*, Comp. Math. **59** (1986), 339–398.

[258] S. Zucker, *L^2-cohomology and intersection homology of locally symmetric varieties III*, Théorie de Hodge, Proc., Luminy 1987, Astérisque, Vols. 179–180 (1989), 245–278.

[259] S. Zucker, *L^2-cohomology of Shimura varieties*, Automorphic Forms, Shimura Varieties, and L-Functions, Proc. Conf. Ann Arbor 1988, Vol. II, Academic Press (1990), 377–391.

CHAPTER 16

Metric Spaces of Curvature $\geqslant k$

Conrad Plaut

Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA
E-mail: cplaut@utk.edu

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

1.1. *Overview*

This article is intended to be an introduction to the theory of metric spaces of curvature bounded below, and a survey of recent results. Time and space constraints have prevented it from being as comprehensive as we originally had planned, but we hope that this article will provide a good beginning point for a student or mathematician interested in this area, and a common reference for future papers on the subject. Through Section 9 we have tried to present good enough sketches of nearly all arguments, so that a dedicated reader can fill in the remaining details without too much difficulty. In Section 9 we provide fewer details, and in the last sections we mostly survey known results.

We do not assume in this article that the reader has any knowledge of Riemannian geometry, although such knowledge is helpful for understanding the motivation behind certain topics. We also occasionally put results into perspective by stating what they mean in terms of Riemannian manifolds, but these parts of the article are not essential. A good working knowledge of topology and basic real analysis should be sufficient background.

Most of the results described in this article were obtained in the current decade. Some of the papers have not been published yet (occasionally due to delays by publishers); some, apparently, will never be published. The field is in a state of rapid development, including simplification of earlier proofs. We have tried to be as inclusive as possible in at least mentioning all closely related (correct) papers of which we were aware when this paper was written in the summer of 1996, and apologize for any omissions.

We have limited ourselves to the subject of spaces of curvature bounded below in the senses of Alexandrov or Berestovskii–Wald, which generalize the notion of bounded sectional curvature in Riemannian manifolds. Although they are included in this class of spaces, we have not specifically discussed the theory of convex surfaces – except by way of historical comments and examples. We have not included results about generalized Ricci curvature (cf. [22]). We include results on the metric geometry of Riemannian manifolds inasmuch as they are also Alexandrov spaces, and we comment on methods from Alexandrov space geometry for attacking problems in Riemannian geometry, but we go no further than this. Good references for metric geometry of Riemannian manifolds are [33] and [35].

Many geometric topologists became interested in questions of convergence of Riemannian manifolds, and the accompanying singular spaces, as a result of the paper of Grove, Petersen, and Wu [43]. One of the main results (Corollary 219 in the present paper) of [43] is a finiteness theorem for a class of Riemannian manifolds. Their proof uses many tools from geometric topology, such as the disjoint disk property and manifold resolutions. The reader may wonder what has happened to such techniques, which do not appear in the present article. Since [43] was written, great progress has been made in understanding the geometric structure of the limit spaces that occur in [43]. The stronger geometric properties that have been discovered have resulted in a smaller reliance on techniques from geometric topology. The main tool for this kind of finiteness theorem is now Perelman's Stability Theorem (Theorem 217), and the topological tools for its proof come entirely from Siebenmann's paper [83]. Nonetheless, we hope that this paper shows that there are other, newer reasons why geometric topologists should remain interested in geometric singular spaces.

In particular, the relationship between geometric topology and metric geometry still seems far from fully understood. For example, the basic question of whether any nonsmooth topological manifolds admit inner metrics of curvature bounded below remains unanswered.

We sincerely apologize for the lack of pictures in this article; time pressure is our only excuse. Thanks are due Valera Berestovskii for providing historical information.

1.2. *Historical comments*

The foundations of the intrinsic geometry of metric spaces lie in Gauss' 1827 work, "General investigations on curved surfaces". In this paper he considers sets S in R^3 having the property that there is a differentiable homeomorphism from an open set in the plane to a neighborhood of any point $p \in S$, whose differential is one-to-one. Such a "regular surface" has an intrinsic geometry in which the distance between two points can be measured as the shortest length of any curve joining those points. Here the length of a curve is measured simply by considering it as a curve in the ambient Euclidean space. On the other hand, the surface has a tangent plane, and the restriction of the ambient inner product to this plane provides a "linear element" which is sufficient to measure the lengths of curves. Gauss' Theorema Egregium states that the Gaussian curvature depends only on this linear element and not on the particular embedding of the surface. That is, curvature is an intrinsic property of the surface, not an extrinsic one. Riemann generalized the notion of regular surface by freeing it entirely of its Euclidean confines: in modern terminology, a Riemannian manifold is a smooth manifold endowed with an inner product on each tangent space (called the *Riemannian metric*), which varies smoothly in a natural sense. One can again measure the lengths of curves by integrating the lengths (in terms of the Riemannian metric) of their tangent vectors, and this measurement in turn gives rise to an intrinsic metric. It is possible to define a notion of *sectional curvature* which generalizes the notion of Gaussian curvature to higher dimensions.

Gauss is also responsible for the first observation that geometry can exert control over topology. The Gauss–Bonnet Theorem, in its simplest global form, states that $\int_S K \, d\sigma = 2\pi \chi(S)$, where K is the Gaussian curvature, the first integral is over the entire surface and $\chi(S)$ is the Euler characteristic of S. In particular, an orientable surface with everywhere positive Gaussian curvature must be homeomorphic to a sphere. One local form of the theorem states $\int_R K \, d\sigma = 2\pi - \sum \theta_i$, where the first integral is over a region R bounded by a triangle whose sides are geodesics (shortest paths) and the sum on the right is of the three exterior angles of the triangle. Note that in a triangle of fixed area in a space of high (positive) curvature, the interior angles of the triangle need to be fairly large: the triangle is "fat", and its perimeter is small relative to its area. With very negative curvature, a triangle of the same area must be very "thin", with larger relative perimeter. In other words, if we fix the side lengths of the triangle, then the above integral, in a positively curved space, integrates a higher curvature over a bigger area; the corresponding interior angles should be larger than in a triangle of same side lengths in a negatively curved space. This observation is the basis for Alexandrov's definition of bounded curvature, which compares the angles of any small triangle in a space with intrinsic metric, to corresponding angles in a "representative" triangle in a "model" surface of constant curvature.

A natural extension of regular surfaces is convex surfaces – the boundaries of convex regions in Euclidean space. A (geometrically) nonsmooth example is the boundary of a standard simplex in $\mathbb{R}^3$. As in the case of a regular surface, the metric in a convex surface is defined by taking the length of the shortest curve (in the surface) joining two points. Convex surfaces generally have no smooth manifold structure naturally compatible with their geometry; clearly they are not always smooth submanifolds of Euclidean space. There are (possibly dense!) "singular points" with no well-defined tangent plane. Nonetheless, such spaces have a well-defined "tangent cone" at every point, consisting of all vectors tangent to shortest paths starting at a point. Following Alexandrov in the 40's [2], much work has been done to understand the structure of convex surfaces – mostly in the Russian geometrical school. The connection to the present article is contained in the following theorem, due to Alexandrov:

THEOREM 1. *A metric space with intrinsic metric is isometric to a convex surface if and only if it is homeomorphic to S^2 and has curvature (in terms of triangle comparisons) $\geqslant 0$.*

Here intrinsic (or inner) metric means that the distance between two points is the infimum of the lengths of curves joining them. Alexandrov's theorem shows that, in same the way that regular surfaces can, using Riemannian geometry, be freed from their embeddings in Euclidean space, convex surfaces can be "abstracted" from their Euclidean confines by considering inner metric spaces having a lower curvature bound. This very large class of spaces, which includes convex surfaces and Riemannian manifolds, as well as their quotients, cones, suspensions, and limits, is the main subject of this article.

It is interesting (given that convex surfaces have a lower curvature bound and not an upper one) that the Russian school initially worked mostly with metric spaces having an upper curvature bound. The primary exception to this trend was work on the question of synthetic differential geometry. This extension of Hilbert's Fifth Problem asks to what degree differential geometry can be developed from basic principles, without assuming the existence of coordinate charts. The problem of synthetic differential geometry was solved by Berestovskii and Nikolaev by 1980 [6,55]. They they showed that a locally compact inner metric space having three metrically defined properties – curvature bounded both above and below, and "geodesic completeness" – are smooth manifolds with $C^{1,\alpha}$ Riemannian metrics (see Theorem 210). In the 80's the interest in convergence of Riemannian manifolds created by Gromov's Precompactness Theorem [32] turned attention towards the singular spaces which can occur as limits of Riemannian manifolds. Quickly it was recognized that if the manifolds have a uniform lower sectional curvature bound, then the limit spaces have a lower curvature bound in the sense of Alexandrov. There followed throughout the 90's an explosion of work on such spaces.

One of the main themes of this article is the tremendous influence that a curvature bound can have over topology, at the infinitesimal, local, and global levels, as well as on a global level across whole families of spaces with geometrically defined restrictions. In this regard modern results in the field remain very close, in spirit, to their earliest ancestor, the Gauss–Bonnet Theorem.

2. Metric fundamentals

Throughout this article, (X, d) will denote a complete metric space. In some sections or subsections we will add further restrictions to X; these will be stated in the first paragraph of the section or subsection in question. We will use the notation d to represent the metric in any metric space, as long as no confusion will result. The *diameter* of X is $\operatorname{diam}(X) = \sup\{d(x, y)\colon x, y \in X\}$. An *isometry* is a distance preserving map; a *local isometry* is a map which is an isometry when restricted to a small enough open set about any point.

DEFINITION 2. For any $\varepsilon > 0$, an ε-net $\mathcal{N}$ in X is a subset of X such that for all $x \in X$ there exists some $y \in \mathcal{N}$ such that $d(x, y) < \varepsilon$.

Since X is complete, it is compact if and only if for every $\varepsilon > 0$ there exists a finite ε-net.

If $c : I \to X$ is a curve, we denote the *length* of c by $L(c) = \sup\{\sum d(x_i, x_{i+1})\}$, where the supremum is taken over all partitions $\{x_1 < \cdots < x_n\}$. By the triangle inequality, $L(c)$ is also the limit as the partitions become fine, of $\sum d(x_i, x_{i+1})$. If $L(c)$ is finite we say that c is *rectifiable*. A rectifiable curve can always be reparameterized in a monotone way so that the parameter is proportional to arclength (cf. [78]). Unless otherwise stated, we will assume that all rectifiable curves are parameterized proportional to arclength. A *unit* curve is one that is parameterized by arclength. Note that any family of unit curves is automatically equicontinuous. We will often use this fact without further comment when applying Ascoli's Theorem.

REMARK 3. Let $\gamma_i : [a, b] \to X$ be arclength-proportional curves in a metric space X, $i = 0, 1, \ldots$. We write $\gamma_i \to \gamma_0$ to denote that γ_i converges uniformly to γ_0. More generally, suppose that the γ_i are not defined on the same interval. Then we write $\gamma_i \to \gamma_0$ if, when each γ_i, $i > 0$, is reparameterized proportional to arclength on the same interval as γ_0, the resulting functions converge uniformly. Note that if $L(\gamma_i)$ is bounded above then $\gamma_i \to \gamma_0$ if and only if γ_i converges to γ_0 pointwise. This situation will occur frequently, and we will use the contents of this remark without further reference.

REMARK 4. When we denote a constant in the form $c(a_1, \ldots, a_k)$ this means that the constant c depends only on the values $a_1, \ldots, a_k$.

2.1. *Inner metric spaces*

Let c be a curve in X joining points x and y. If $d(x, y) = L(c)$ then c is called *minimal*. Henceforth the notation γ_{ab} will always denote a minimal curve joining a and b. A curve $\gamma : I \to X$ is called a *geodesic* if for each interior value t of I, γ is minimal when restricted to a small interval centered at t. The metric of X is called *inner* (or *intrinsic* or *length*) if for all $x, y \in X$, $d(x, y) = \inf\{L(c)\colon c \text{ joins } x \text{ and } y\}$. If each pair of points in X can be joined by a minimal curve, X is sometimes called a *geodesic space*.

Inner metric spaces can be considered as geometric objects, as opposed to merely topological or analytical objects. For example, one can consider the subspace metric on the unit sphere in Euclidean space. This metric is compatible with the subspace topology, and so provides an adequate rough measure of "closeness" between points. However, this metric is inadequate for describing the internal geometry of the sphere. It is better to consider the *inner metric induced by* the subspace metric – i.e., the distance between two points is the length of the shortest curve joining them. One then obtains the metric of the standard geometry of the sphere with Gaussian curvature 1. This geometry describes the sphere as it would appear to a creature confined to live in the surface of the sphere, who, for example, would measure the distance between antipodal points as π, not 2.

DEFINITION 5. For any $x, y, z \in X$, we define the *excess* of the triple by $\varepsilon(y; x, z) := d(x, y) + d(y, z) - d(x, z)$. If $\varepsilon(y; x, z) = 0$ and x, y, z are distinct, we say that y is *between* x and z; in this case, if $d(x, y) = d(y, z)$, then we say that y is a *midpoint* of x and y. We define the *strong excess* of the triple to be

$$\sigma(y; x, z) = \frac{\varepsilon(y; x, z)}{\min\{d(x, y), d(y, z)\}}.$$

The next lemma is an immediate consequence of the definition and the triangle inequality.

LEMMA 6. *Let w, x, y, z be points in X. Then*
(1) $\varepsilon(z; y, w) + \varepsilon(y; x, w) = \varepsilon(y; z, x) + \varepsilon(z; w, x)$, *and*
(2) $\varepsilon(z; y, w) \leqslant \varepsilon(y; z, x) + \varepsilon(z; w, x)$.

PROPOSITION 7. *X is an inner metric space if and only if for every $x, z \in X$ and $\varepsilon > 0$ there exists a $y \in X$ such that $d(x, y) = d(x, z)/2$ and $\varepsilon(y; x, z) < \varepsilon$.*

A point y as in the proposition will be called an "almost midpoint" between x and z. To prove the proposition, note that if X is an inner metric space, then by choosing a curve c joining of length sufficiently close to $d(x, z)$, then the point y which bisects c is an almost midpoint. Conversely, we suppose, for simplicity, that $d(x, z) = 1$. We choose a point $\phi(1/2)$ such that

$$d\left(x, \phi\left(\frac{1}{2}\right)\right), d\left(z, \phi\left(\frac{1}{2}\right)\right) \leqslant \frac{1}{2} + \frac{\varepsilon}{4}.$$

Continuing in this way, we can construct a mapping ϕ from the dyadic rationals in $[0, 1]$ into X with the property that if, for dyadic rationals w, y, $d(w, y) = 2^{-n}$, then

$$d(\phi(w), \phi(y)) \leqslant 2^{-n} + \frac{2^n - 1}{4^n}\varepsilon < (1 + \varepsilon) d(w, y).$$

Clearly ϕ can be extended (using the completeness of X) to all of $[0, 1]$ to produce a curve joining x and z whose length is less than $1 - \varepsilon$.

In a metric space in general, the closure of an open metric ball $B(p,r)=\{x\colon d(x,p)<r\}$ is contained in, but not equal to the closed metric ball $\{q\colon d(p,q)\leqslant r\}$. However, equality is easily seen to be true in an inner metric space. For this reason, we will use the notation $\overline{B}(p,r)$ to denote $\{q\colon d(p,q)\leqslant r\}$ in an inner metric space.

The following is a theorem of Cohn-Vossen [26].

THEOREM 8. *An inner metric space X is locally compact if and only if every closed, bounded set is compact. Every pair of points in a locally compact inner metric space can be joined by a minimal curve.*

The hardest part of the theorem is showing that local compactness implies that closed metric balls are compact. To do this, one can show that, for any p, if $W=\{r\colon \overline{B}(p,r)$ is compact$\}$ then $\sup W=\infty$. Here the main step is showing that if $s\in W$ for all $s<r$ then $r\in W$. Given any sequence $\{a_i\}$ in $\overline{B}(p,r)$ one can use the inner metric to find, for any n, points b_i^n such that $b_i^n\in\overline{B}(p,r-2^{-n})$ and $d(a_i,b_i^n)\leqslant 2^{-n+1}$. Since we can extract a convergent subsequence of $\{b_i^n\}$, we can assume, by taking a subsequence, that $d(a_i,a_j)\leqslant 2^{-n+3}$. The proof is completed using a standard diagonal argument. Minimal curves can now be constructed using Ascoli's Theorem.

Any "reasonable" metric space (X,d) has an induced inner metric d_I (see Definition 65 and subsequent comments).

One property of inner metrics that distinguishes them from ordinary metrics is that they need only be defined locally. Indeed, inner metrics can be defined by merely producing a reasonable way to measure the lengths of sufficiently many curves in a neighborhood of every point – as is the case with Riemannian manifolds (see Example 94). We illustrate this point with the following proposition:

PROPOSITION 9. *Let $\pi: X\to Y$ be a covering map and suppose d is an inner metric on Y. Then if X is arcwise connected, there exists a unique inner metric $\tilde{d}$ on X such that π is a local isometry.*

The metric $\tilde{d}$, which is called the *lift* of d, is defined in the following way: for any pair of points $x,y\in X$, $\tilde{d}(x,y)=\inf\{L(\pi(c))\}$, where the infimum is over all curves c joining x and y in X. In essence we are measuring the length of a curve in X by projecting it and measuring the length of the resulting curve in Y. This proposition, as well as the following corollary, are false for metric spaces in general.

COROLLARY 10. *If X and Y are inner metric spaces and $\phi: X\to Y$ is a one-to-one local isometry, then ϕ is an isometry.*

2.2. *Space forms and cosine laws*

We denote by S_k^n the n-dimensional simply connected space form (sphere, Euclidean space, or hyperbolic space) of curvature k. When $n=2$ we will omit the n. The spaces S_k^n are metrically characterized in the following way. First, S_k^n is a locally compact inner metric space. For any unit geodesics γ_1, γ_2 starting at a point, the quantity

$$\alpha(\gamma_1, \gamma_2) := 2 \lim_{t \to 0} \sin^{-1} \frac{d(\gamma_1(t), \gamma_2(t))}{2t}$$

exists and satisfies the following *cosine laws* for all s, t:

$$\cos\alpha(\gamma_1, \gamma_2) = \frac{\cos\sqrt{k}d(\gamma_1(s), \gamma_2(t)) - \cos\sqrt{k}s\cos\sqrt{k}t}{\sin\sqrt{k}s\sin\sqrt{k}t} \quad \text{for } k > 0, \tag{1}$$

$$\cos\alpha(\gamma_1, \gamma_2) = \frac{s^2 + t^2 - d(\gamma_1(s), \gamma_2(t))^2}{2st} \quad \text{for } k = 0, \tag{2}$$

$$\cos\alpha(\gamma_1, \gamma_2) = \frac{\cosh\sqrt{|k|}d(\gamma_1(s), \gamma_2(t)) - \cosh\sqrt{|k|}s\cosh\sqrt{|k|}t}{\sinh\sqrt{|k|}s\sinh\sqrt{|k|}t} \quad \text{for } k < 0. \tag{3}$$

Note that, with the usual relationship between trigonometric and hyperbolic functions (e.g., $\cosh x = \cos ix$), the cosine laws for $k < 0$ and $k > 0$ are essentially the same. Furthermore, if we take the limit as $k \to 0$, these formulas approach the cosine law for $k = 0$. The formula can be readily seen for $k > 0$ from $\lim_{x\to 0} \frac{\sin x}{x} = 1$ and the following reformulation of the spherical cosine law [78]:

$$\sin^2\left(\frac{\alpha_k(\gamma_1, \gamma_2)}{2}\right) = \frac{\sin\frac{1}{2}\sqrt{k}(d(\gamma_1(s), \gamma_2(t)) + s - t)\sin\frac{1}{2}\sqrt{k}(d(\gamma_1(s), \gamma_2(t)) - s + t)}{\sin\sqrt{k}s\sin\sqrt{k}t}. \tag{4}$$

It is also evident from formula (4) (and its hyperbolic counterpart) that, if k is fixed and s or t gets small, the right-hand side of formula (4) tends to the Euclidean formula. In other words, the space forms are "infinitesimally" Euclidean.

The metric characterization is not the most useful way to understand the space forms; however, one of our goals is to minimize the knowledge of differential geometry that the reader must have. Therefore we confine our comments to only the most general ones, with no proofs given. For more details, consult one of the books on differential geometry in the bibliography. For the rest of this paragraph assume $k > 0$. The sphere S_k^n is the sphere of radius $1/\sqrt{k}$ in $\mathbb{R}^{n+1}$, with the Riemannian metric induced by the inclusion. Equivalently, we can measure the distance between two points by the length of the shortest smooth (equivalently piecewise smooth) curve joining them, where the length is measured in the usual way for smooth curves in Euclidean space. With this distance, the sphere is isometric to the one given by the cosine law above. Being a smooth submanifold of $\mathbb{R}^{n+1}$, S_k^n has a tangent space at every point. Tangent to each unit vector is a unique great circle, which is a geodesic in both the metric and differential geometric senses. The angle between two tangent vectors (measured using the Euclidean inner product on the tangent space) is the same as the (metric) angle defined above, between the corresponding geodesics. The great circles are the only geodesics, and are minimizing up to length $\pi/\sqrt{k}$, which is the (intrinsic!) diameter of S_k^n. We can therefore refer to the *space of directions* at a point p, which will mean both the unit sphere in the tangent space at p, and equivalently the space of unit geodesics starting at p with the angle as metric. In either case, the space of directions is isometric to S_1^{n-1}. Finally, note that every S_k^n can be obtained from the unit sphere S_1^n by scaling: multiplying the distance function by the factor $1/\sqrt{k}$. When scaling the metric, geodesics clearly remain the same. In addition, it follows from the fact, mentioned above, that the space forms are infinitesimally Euclidean, that the angle between geodesics is not changed by scaling. In other words, scaling really just changes the size, not the essential geometry, of the space form.

As in the case of the sphere, if $k < 0$, S_k^n can be obtained from the hyperbolic space S_{-1}^n by scaling by the factor $1/\sqrt{-k}$. Hyperbolic space is homeomorphic to Euclidean space. Unit geodesics starting at a point with the angle metric form a space of directions homeomorphic to S_1^{n-1}. The geodesics starting from a point spread out much more rapidly than geodesics of the same angle in Euclidean space; consequently geodesics in S_{-1}^n, like those in Euclidean space but unlike those in the sphere, never come back together again. Also, every segment of a geodesic in S_{-1}^n is minimizing (as in Euclidean space, but not the sphere). In the space forms it should now be clear that, the lower the curvature, the more rapidly geodesics diverge. Later we will see that measuring the rate of divergence of geodesics is an excellent way to understand curvature.

DEFINITION 11. X is said to be *(metrically) n-point homogeneous* if for every $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_n\} \subset X$ such that $d(x_i, x_j) = d(y_i, y_j)$ for all i, j, there is an isometry $\phi : X \to X$ such that $\phi(x_i) = y_i$ for all i. A 1-point homogeneous space is simply called *homogeneous*.

Using Cartan's Theorem (cf. [27]), the following can be shown by induction on m.

PROPOSITION 12. *S_k^n is m-point homogeneous for all m.*

2.3. *Distance geometry, monotonicity*

Let M be a finite metric space. A basic question in distance geometry is when M can be embedded in a space form. If $M = \{x, y, z\}$ has three points, it is not hard to see geometrically that M can be (isometrically) embedded in any space form S_k so long as $d(x, y) + d(x, z) + d(y, z) \leqslant 2\pi/\sqrt{k}$. When $k \leqslant 0$ we always take $k^{-1/2} = \infty$ (e.g., when considering radii of metric balls). The image of such an embedding (which is unique up to isometric motion of S_k, by Proposition 12) gives rise to a *representative triangle* $T = T(\tilde{x}, \tilde{y}, \tilde{z})$ in S_k.

DEFINITION 13. A *triple* $(y; x, z)$ is subset of three points in a metric space. The distinguished point y is the *vertex* of the triple. We denote by $\alpha_k(y; x, z) \in [0, \pi]$ the angle of a representative triangle in S_k at the point corresponding to y.

As can be seen from the cosine laws and the results of the previous section, the function $\alpha_k(y; x, z)$ is a continuous function of k, $d_1 := d(x,z)$, $d_2 := d(y,x)$ and $d_3 = d(y,z)$, and is monotone increasing in either of the first two variables (leaving the other variables fixed). Also, $\alpha_k(y; x, z) = \pi$ if and only if y is between x and z, and $\alpha_k(y; x, z) = 0$ if and only if one of x, z is between y and the other. In general, $\alpha_k(y; x, z)$ is defined for all k such that $d(x,y) + d(x,z) + d(y,z) \leqslant 2\pi/\sqrt{k}$.

REMARK 14. We will refer to the monotonicity of $\alpha_k(x; y, z)$ in k and d_1 as the *first monotonicity* and *second monotonicity*, respectively. If $d(x,y) + d(x,z) + d(y,z) > 2\pi/\sqrt{k}$ then we take $\alpha_k(y; x, z) = \infty$.

There is a close relationship between $\sigma(b; a, c)$ and $\alpha_k(b; a, c)$. The following lemma is evident from the Cosine Laws or elementary Riemannian geometry:

LEMMA 15. *Let k be fixed. Then $\sigma(b; a, c)$ is small if and only if $\alpha_k(b; a, c)$ is close to π.*

More precisely, for every $\varepsilon > 0$ there exist constants $\delta_1(k), \delta_2(k) > 0$ such that if $\sigma(b; a, c) < \delta_1$ then $\alpha_k(b; a, c) > \pi - \varepsilon$ and if $\alpha_k(b; a, c) > \pi - \delta_2$ then $\sigma(b; a, c) < \varepsilon$. Often we are interested in showing that $\alpha_k(b; a, c) \to \pi$ as a, b, c converge to some point. We can use the following criterion:

LEMMA 16. *Let $M \geqslant 1$ and k be fixed. Then for every $\varepsilon > 0$ there exists a $\delta(k, M) > 0$ such that if $d(a,b) \leqslant M$ and $\frac{1}{M} \leqslant \frac{d(a,b)}{d(b,c)} \leqslant M^2$ and $\varepsilon(b; a, c) \leqslant \delta \cdot (d(a,b) + d(b,c) + d(a,c))$ then $\alpha_k(b; a, c) \geqslant \pi - \varepsilon$.*

To prove the lemma, observe that by the first monotonicity it is sufficient to prove it for $k < 0$. If $d(a,b) = M$, then $d(a,c) \geqslant 1/M^2$ and the statement follows easily from the cosine laws or a compactness argument. The proof is now complete by observing that, if T is a triangle in S_k ($k < 0$) with side lengths S_1, S_2, S_3, and T' is another triangle in S_k with side lengths mS_1, mS_2, mS_3 for some $m \geqslant 1$, then the angles in T are all larger than the corresponding angles in T'. (This fact follows from the first monotonicity and elementary Euclidean geometry.) Thus by "enlarging" a given triangle until $d(a,b) = M$, we are finished.

Later we will be using representative angles to approximate angles (which will be defined later). The following facts are convenient. First, we say that two sequences $\{(x_{i2}; x_{i1}, x_{i3})\}$ and $\{(y_{i2}; y_{i1}, y_{i3})\}$ are *(metrically) equivalent* if

$$\lim_{i\to\infty} \frac{|d(x_{ij}, x_{ik}) - d(y_{ij}, y_{ik})|}{\min_{j,k}\{d(x_{ij}, x_{ik})\}} = 0 \quad \text{for all } j, k.$$

PROPOSITION 17. *Let $\{(b_i; a_i, c_i)\}$ be a sequence of triples such that $\alpha_k(i) := \alpha_k(b_i; a_i, c_i)$ is defined for all i. Then*

(1) *If $\lim_{i\to\infty} \frac{d(a_i, c_i)}{\min\{d(a_i, b_i), d(b_i, c_i)\}} = 0$, $\alpha_k(i) \to 0$.*

(2) *If $\{(b_i'; a_i', c_i')\}$ is a sequence of triples equivalent to $\{(b_i; a_i, c_i)\}$ such that $\alpha_k'(i) = \alpha_k(b_i'; a_i', c_i')$ is defined for all i, then $\lim_{i\to\infty} \alpha_k(i)$ exists if and only if $\lim_{i\to\infty} \alpha_k'(i)$ exists, and if both exist they are equal.*

The first part of the proposition can be proved using the fact that the space forms are infinitesimally Euclidean (cf. the comments after formula (4)). To prove the second part, note first that the argument can be reduced to the case when the sets of triples have two "side lengths" in common, by replacing $(b_i; a_i, c_i)$ with a triple having one side length in common with $(b'_i; a'_i, c'_i)$, then replacing it with one having two sides in common. But if the triples have two sides in common, then the second part follows from the first part and the triangle inequality.

It is not possible, generally, to embed a given four point metric space in *any* space form. For example, the quadruple $M = \{a, b, c, d\}$ such that $d(a, b) = d(b, c) = d(b, d) = 1$ and $d(a, c) = d(a, d) = 2$ is *branching*; that is, there is a point (b) that is simultaneously between two other pairs of points ((a, c) and (a, d)), but M cannot be embedded in the real line. If this quadruple were embedded as a quadruple $\{A, B, C, D\}$ in a space form, the minimal geodesics joining A to C and A to D would coincide as a minimal geodesic from A to B; that is, the geodesic from A to B would "branch" at B. Since geodesics in a space form (or any Riemannian manifold!) never branch, M cannot be embedded in any S^n_k. A quadruple M will be said to be *linear* if it can be embedded in the real line, and *nondegenerate* if no point is between two other points. If a quadruple $M = \{a, b, c, d\}$ can be embedded in S^3_k, then we say that k is an *embedding curvature* for M. If M can be embedded in S_k, we say k is a *planar* embedding curvature for M. Note that M can be embedded in S^3_k if and only if M can be embedded in S^n_k for all $n > 3$. In [10] the following theorem is proved:

THEOREM 18. *Let M be a (metric) quadruple. Then exactly one of the following is true:*

(1) *M is nondegenerate and the embedding curvatures are an interval of the form $[K_1, K_2]$ or $(-\infty, K]$, and the endpoints of these intervals are precisely the planar embedding curvatures for M.*
(2) *M is branching and so cannot embed in any space form.*
(3) *M is linear.*
(4) *M has a unique embedding curvature k, which is planar.*

We will give a sketch of the proof of the theorem, for which we need the following notation:

DEFINITION 19. Let $M = \{x_1, x_2, x_3, x_4\}$ be a (metric) quadruple. For any k, and i, j, m, n distinct, we let

$$V_k(x_i) = \alpha_k(x_i; x_j, x_m) + \alpha_k(x_i; x_j, x_n) + \alpha_k(x_i; x_m, x_n),$$

$$A_k(M) = \max_i \{V_k(x_i)\}.$$

From the properties of α_k, we see that $V_k(x_i)$ is also continuous and monotone increasing in k. Now suppose that $M = \{a, b, c, d\}$ is isometrically embedded in S^3_k and the image of an embedding is the points A, B, C, D. Then the geodesics from A to the other three points are directions $\gamma_B, \gamma_C, \gamma_D$ at A, that is, they form a triple of points embedded in the space of directions at A, which is S_1. The converse is also true: if $\alpha_k(a; b, c), \alpha_k(a; c, d), \alpha_k(a; b, d)$

are the distances for some triple in S_1 then M embeds in S_k^3. From what we have already seen about embedding triples of points in space forms, we immediately get the following proposition. (The nondegeneracy is needed to ensure none of the angles is 0.)

PROPOSITION 20. *A nondegenerate quadruple $M = \{x_1, x_2, x_2, x_4\}$ embeds in S_k^3 if and only if $A_k(M) \leqslant 2\pi$ and the triangle inequality holds for any three values $\alpha_k(x_i; x_j, x_m)$, $\alpha_k(x_i; x_j, x_n)$, $\alpha_k(x_i; x_m, x_n)$. Furthermore, if $\alpha_k(x_i; x_j, x_m) = \alpha_k(x_i; x_j, x_n) + \alpha_k(x_i; x_m, x_n)$ for some i then the embedding is planar.*

Now, if for some $a \in M$, $V_k(a) = 2\pi$ then each angle is $\leqslant \pi$ but the sum of any two is $\geqslant \pi$, so the triangle inequality is always true. M can be embedded in S_k^3, but more exactly, the three corresponding directions span a 2-plane in the tangent space, so we have in fact embedded M in S_k. It is shown in [10, Lemma 3], that if K is the largest number for which $A_K(M)$ is defined then $A_K(M) \geqslant 2\pi$. On the other hand, the properties of α_k indicate that for some $k < K$, $A_k(M) < 2\pi$. By continuity there must be an m such that $A_m(M) = 2\pi$. Thus we have a planar embedding of M in S_m, and m is the maximal possible value for which M embeds in S_m^3. It is also shown in [10, Lemma 4], that if there is another planar embedding, i.e., an embedding of M in S_h, with $h < m$ then there are no embeddings of M in S_k^3 for any $k < h$. We see that there are at least one and at most two planar embeddings, and they are extremal. In light of Proposition 20 and the above comments, we see that M has a nonplanar embedding in S_k^3 if and only if $A_k(M) < 2\pi$ and the triangle inequality is satisfied and strict for all triples of angles. Therefore the set of all such k is open, and its boundary consists of values m such that M embeds in S_m. The remainder of the proof of Theorem 18 is a case-by-case analysis. A similar case-by-case analysis can be used to prove the following, which is not stated in [10].

COROLLARY 21. *Let M be a quadruple. Then M has a planar embedding number $m \geqslant k$ if and only if $A_k(M) \leqslant 2\pi$.*

3. Curvature bounded below

3.1. *Wald–Berestovskii curvature $\geqslant k$*

The following definition is due to Berestovskii [10], and is a modification of one due to Wald [88].

DEFINITION 22. X is said to have curvature $\geqslant k$ if for all $x \in X$ there is an open set $U \subset X$ containing x such that every quadruple of points $\{a, b, c, d\}$ in U isometrically embeds (as a metric space with the induced metric) in S_m for some $m \geqslant k$. Such a set U will be called a *region of curvature $\geqslant k$*.

A natural question to ask is: Why four points? Certainly the results of the previous section show that three points is too few. On the other hand, Berestovskii has shown [10] that if one replaces quadruples by quintuples in the above definition, then one gets essentially only spaces of constant curvature. So only quadruples allow the possibility of an interesting theory beyond the geometry of space forms. As an immediate consequence of Corollary 21 we get the following formulation of bounded curvature, which has the advantage of involving only the single model space S_k.

PROPOSITION 23. *An open set U in X is a region of curvature $\geqslant k$ if and only if every quadruple $\{a, b, c, d\}$ in U satisfies $V_k(a) \leqslant 2\pi$.*

It is well-known that every separable compact metrizable space embeds topologically in the Hilbert Cube $\mathrm{H} = I_1 \times I_2 \times \cdots$, where $I_n = [0, 2^{-n}]$. Since H itself is naturally a convex subset of the (geometrically Euclidean) separable Hilbert space, it has curvature $\geqslant 0$ with the induced metric (in fact it is "flat" in some sense). It follows that every compact metrizable space admits a metric of nonnegative curvature. We see that simply having such a metric imposes no topological restrictions on the space. As we have mentioned earlier, it is also true that simply having an inner metric imposes few restrictions on a metric space. It is therefore somewhat remarkable that the combination of these two conditions imposes, as we will see, a very strong geometric and topological structure.

3.2. *The extended Hopf–Rinow theorem*

DEFINITION 24. Let γ_{ab} be a minimal curve in X. The curve γ is called *extendable* beyond b if γ is the restriction of a minimal curve having b in its interior, and *almost extendable* beyond b if for every $\varepsilon > 0$ there exists a $c \in X\backslash\{a, b\}$ such that $\sigma(b; a, c) < \varepsilon$.

Note that Definition 24 is stronger than the definition of "almost extendable" given in [76]. It easy to see that an extendable minimal curve is almost extendable.

DEFINITION 25. An inner metric space with the property that every minimal curve γ_{ab} has a restriction that is extendable beyond b is called *geodesically complete.*

Geodesic completeness is equivalent to the maximal domain of definition of every geodesic being all of $\mathbb{R}$. The Hopf–Rinow Theorem in Riemannian geometry states, for a Riemannian manifold M, the equivalence of metric completeness and geodesic completeness, and, if M is complete in this way, gives the existence of minimal curves between all pairs of points. In a Riemannian manifold, locally, joins by minimal curves are unique. We will see that geodesic completeness generally fails for spaces of curvature bounded below (and can fail badly even for 2-dimensional spaces). Almost extendability of geodesics serves as its replacement.

DEFINITION 26. For any $p \in X$ let J_p be the set of all $q \in X$ such that there exists a unique minimal curve γ_{pq} which is almost extendable beyond q.

THEOREM 27. *If X is an inner metric space of curvature $\geqslant k$ then J_p contains a dense G_δ subset.*

We call the above theorem an "extended" Hopf–Rinow Theorem, even though the terminology is somewhat misleading. Even for Riemannian manifolds, this theorem does not follow from the Hopf–Rinow Theorem: being geodesically complete does *not* mean that every minimal curve is extendable in the sense of Definition 24. (A Riemannian manifold with the latter property must be diffeomorphic to its tangent space via the exponential map. Theorem 27 also includes information about the cut locus of the manifold.)

The proof of the theorem needs some preliminaries.

PROPOSITION 28. *Let X be an inner metric space and suppose $p, q \in X$ have the property that $B := B(q, r)$ is a region of curvature $\geqslant k$ for some $r > 0$, and there exist q_i with $q_i \to q$ and $\sigma(q; q_i, p) \to 0$. For any positive $t < \min\{r, d(p, q)\}$ and $\delta > 0$ there exists an $\varepsilon > 0$ such that if $a_j \in B$, $j = 1, 2$, satisfy $d(q, a_j) = t$ and $\varepsilon(a_j; p, q) < \varepsilon$, then $d(a_1, a_2) < \delta$.*

PROOF. Suppose $d(q, a_i) = t$. The proposition will follow from Lemma 15 if we can show that for any $\eta > 0$ there exists an $\zeta > 0$ and i large enough that if $\varepsilon(a_j; p, q) < \zeta$ then $\sigma(q; a_j, q_i) < \eta$. For then we know that $\alpha_k(q; a_j, q_i)$ is close to π, and by Proposition 23 $\alpha_k(q; a_1, a_2)$ is small, and by Proposition 17 $d(a_1, a_2)$ is small. Choose i so that $\sigma(q; q_i, p) \leqslant \eta/2$ and $d(q, q_i) < t$. By Lemma 6, if

$$\varepsilon(a_j; p, q) < \zeta, \qquad \varepsilon(q; a_j, q_i) < \zeta + \frac{\eta d(q, q_i)}{2},$$

so we need only choose $\zeta < \eta d(q, q_i)/2$. □

We return to the proof of Theorem 27. Let L_i be the set of all $y \in X$ such that for some z, $\sigma(y; p, z), d(y, z) < 2^{-i}$, and let $J'_p = \bigcap_{i=1}^{\infty} L_i$. Obviously J'_p is a G_δ. The proof that J'_p is dense requires only properties of the inner metric (this fact allows us to get the global conclusion even though the curvature assumption is only local). Basically, we start with a constant $\rho_0 > 0$, begin at an arbitrary point x, and choose a point x_1 such that $d(x_1, x) = \rho_0/2$ and $\sigma(x_1; p, x) < 1/2$. Continuity of the distance function implies that there is a $\rho_1 < \rho_0/4$ such that if some point y satisfies $d(y, x_1) \leqslant \rho_1$ then $\sigma(y; p, x) < 1/2$. Now we choose x_2 such that $d(x_2, x_1) = \rho_1$ and $\sigma(x_2; p, x_1) < 1/4$. Continuing this process we can construct a Cauchy sequence $\{x_i\}$ such that $z := \lim x_i$ is in J'_p and $d(x, z) \leqslant \rho_0$.

To complete the proof we need to construct a minimal curve from p to any $q \in J_p$. As in the usual proof of the Hopf–Rinow Theorem in Riemannian geometry, we start at q and construct our curve "backwards" towards p. We suppose that $B(q, r)$ is a region of curvature $\geqslant k$. Let $t < r$ be positive and choose points x_i such that $d(x_i, q) = t$ and $\varepsilon(x_i; p, q) \to 0$. (The points x_i can be found by choosing almost minimal curves from p to q and using the intermediate value theorem to get $d(x_i, q) = t$.) Then Proposition 28 implies that the sequence $\{x_i\}$ is Cauchy, and converges to a point $\gamma(t)$ between p and q such that $d(\gamma(t), q) = t$. Doing this for every $t \in (0, r)$ and letting $\gamma(0) = q$ we obtain a geodesic $\gamma : [0, r) \to X$ starting at q such that the restriction of γ to any $[0, t)$ with $t < r$ is minimal. Then $\lim_{t \to r} \gamma(t)$ exists; we call it $\gamma(r)$. By continuity, $\gamma : [0, r] \to X$ is a minimal curve and $\gamma(r)$ is between p and q. For any $t_i \to r$, $t_i < r$, $\varepsilon(\gamma(r); p, \gamma(t_i)) = 0$ and $\gamma(t_i) \to \gamma(r)$, so $\gamma(r)$ satisfies the hypothesis of Proposition 28. Now we can continue extending γ beyond $\gamma(r)$ as above. By a standard open-and-closed argument we eventually reach p.

The above argument also proves the following corollary, which will be true globally once we have obtained the Global Comparison Theorem (Theorem 43).

COROLLARY 29. *Let X be an inner metric space and suppose $p \in X$ has the property that $B := B(p,r)$ is a region of curvature $\geqslant k$. Then $q \in J_p \cap B$ if and only if there exist points q_i with $q_i \to q$ and $\sigma(q; q_i, p) \to 0$.*

By the Baire Category Theorem, the intersection of countably many dense G_δ sets is again a dense G_δ. We immediately have the following:

COROLLARY 30. *If X has curvature $\geqslant k$ and $x_1, x_2, \ldots \in X$, there exist points x_i' arbitrarily close to x_i such that every pair x_i', x_j' is joined by a unique minimal curve. Furthermore, we can take $x_1 = x_1'$.*

For example, if we are given three points x, y, z in X, we can find points y', z' arbitrarily close to y, z, respectively, such that x, y', z', form the corners of a triangle of minimal curves in X.

Note that if γ is a minimal curve from y to x, any point arbitrarily close to x in the interior of γ is in J_y.

3.3. *Alexandrov's comparisons*

In this subsection X will always be assumed to be an inner metric space of curvature $\geqslant k$, and U will denote a region of curvature $\geqslant k$. A *triangle* in X consists of three minimal curves $T = (\gamma_{ab}, \gamma_{ac}, \gamma_{bc})$. We allow "degenerate" triangles, in which the geodesics may coincide in part. A *hinge* is a pair of minimal curves $(\gamma_{ab}, \gamma_{ac})$. *We will always assume that* γ_{ab} and γ_{ac} are parameterized by arclength on $\mathbb{R}^+$ so that $\gamma_{ab}(0) = \gamma_{ac}(0) = a$. If d lies on γ_{ac} and $\gamma_{ac}, \gamma_{ab} \subset U$, then we can embed a, b, c, d in S_m, for some $m \geqslant k$; the image of this quadruple consists of the corners of a representative triangle $T(\tilde{a}, \tilde{b}, \tilde{c})$ (having the same side lengths) in S_m, with a point $\tilde{d}$ on $\gamma_{\tilde{a}\tilde{c}}$. If we consider the representative triangle $T(\tilde{a}', \tilde{b}', \tilde{c}')$ in S_k and the point $\tilde{d}'$ on $\gamma_{\tilde{a}'\tilde{c}'}$ such that $d(\tilde{a}', \tilde{d}') = d(a, d)$, then by the first monotonicity (see Section 2.2), $d(b, d) \geqslant d(\tilde{b}', \tilde{d}')$. We refer to the points $\tilde{d}$ and $\tilde{d}'$ as the *point corresponding to* d in S_m, S_k, respectively. They are uniquely determined by the representative triangle. Choosing a point e on γ_{ab} and applying the above argument to the hinge $(\gamma_{ab}, \gamma_{ad})$ and the point e, together with the second monotonicity, we prove the following statement, which is one of Alexandrov's curvature conditions.

A0 Given any triangle $T = (\gamma_{ab}, \gamma_{bc}, \gamma_{ca})$ in U, $d \in \gamma_{ac}$, and $e \in \gamma_{ab}$, there exists a representative triangle $\tilde{T}$ in S_k. If $\tilde{d}, \tilde{e}$ are the points on $\tilde{T}$ corresponding to d, e, respectively, then $d(d, e) \geqslant d(\tilde{d}, \tilde{e})$.

REMARK 31. One can consider an a priori weaker condition than A0 by fixing $e = b$ in the statement of A0 and only allowing d to vary. It is a nice geometric exercise using the second monotonicity to show that this "one-point" version of A0 is in fact equivalent to the above version. We will also refer to the one-point version as A0; often it is easier to verify.

An immediate consequence of A0 and the second monotonicity is that for any hinge $(\gamma_{ab}, \gamma_{ac})$, the function $\alpha(s, t) = \alpha_k(a; \gamma_{ab}(s), \gamma_{ac}(t))$ is monotone decreasing in each variable, into $[0, \pi]$. Thus the following definition makes sense.

DEFINITION 32. Given a hinge $(\gamma_{ab}, \gamma_{ac})$, with $a \in U$, the *angle* of the hinge in X is the number

$$\alpha(\gamma_{ab}, \gamma_{ac}) = \lim_{s,t\to 0} \alpha_k\bigl(a; \gamma_{ab}(s), \gamma_{ac}(t)\bigr) = \lim_{s,t\to 0} \alpha_m\bigl(a; \gamma_{ab}(s), \gamma_{ac}(t)\bigr) \in [0, \pi]$$

for any fixed m.

The fact that the angle is independent of k follows from formula (4); in fact, we could even let m vary, as long as it stays bounded. Thus, we can measure angles, when convenient, using representatives in any space form, including the Euclidean plane. We summarize this property in the following convenient form:

LEMMA 33. *Let γ, β be unit geodesics starting at $p \in U$. Then*

$$d\bigl(\gamma(t), \beta(t)\bigr) = 2t \sin\frac{\alpha(\gamma, \beta)}{2} + \mathrm{o}(t).$$

The following basic result is the triangle inequality for angles, and is a consequence of the triangle inequality for the metric of X, and Definition 32:

PROPOSITION 34. *For any three minimal curves $\gamma_{ab}, \gamma_{ac}, \gamma_{ad}$ in X,*

$$\alpha(\gamma_{ab}, \gamma_{ad}) \leqslant \alpha(\gamma_{ab}, \gamma_{ac}) + \alpha(\gamma_{ac}, \gamma_{ad}).$$

The next two statements, which are also curvature conditions of Alexandrov, follow from A0 and are equivalent by the second monotonicity. Note that our meaning of angle coincides with the usual one if X is a Riemannian manifold. Given a hinge $H = (\gamma_{ab}, \gamma_{ac})$ in X, a *representative* of H in S_k consists of a hinge $\tilde{H} = (\gamma_{\tilde{a}\tilde{b}}, \gamma_{\tilde{a}\tilde{c}})$ in S_k such that $d(a, b) = d(\tilde{a}, \tilde{b})$, $d(a, c) = d(\tilde{a}, \tilde{c})$ and $\alpha(\gamma_{ab}, \gamma_{ac}) = \alpha(\gamma_{\tilde{a}\tilde{b}}, \gamma_{\tilde{a}\tilde{c}})$.

A1 For any triangle $T = (\gamma_{ab}, \gamma_{ac}, \gamma_{bc})$ in U, if $\widetilde{T} = (\gamma_{\tilde{a}\tilde{b}}, \gamma_{\tilde{a}\tilde{c}}, \gamma_{\tilde{b}\tilde{c}})$ is a representative of T in S_k, then $\alpha(\gamma_{ab}, \gamma_{ac}) \geqslant \alpha(\gamma_{\tilde{a}\tilde{b}}, \gamma_{\tilde{a}\tilde{c}})$.

A2 For any hinge $H = (\gamma_{ab}, \gamma_{ac})$ in U, there exists a representative $\widetilde{H} = (\gamma_{\tilde{a}\tilde{b}}, \gamma_{\tilde{a}\tilde{c}})$ of H in S_k, and $d(b, c) \leqslant d(\tilde{b}, \tilde{c})$.

It is important to consider when equality occurs in conditions A1 and A2, which we will denote by EA1 and EA2 (cf. Appendix to [36]). In the locally compact case, suppose we have a triangle $T = (\gamma_{ab}, \gamma_{ac}, \gamma_{bc})$ which satisfies EA1 and is contained in a much larger region of curvature $\geqslant k$. Then applying A0 and A2, we see that the distance between any pair of points on γ_{ab} and γ_{ac}, respectively, must be equal to the corresponding distance on a representative triangle. For any d on γ_{bc}, and minimal curve γ_{ad}, applying both Proposition 34 and A1, we see that both $(\gamma_{ab}, \gamma_{ad})$ and $(\gamma_{ad}, \gamma_{ac})$ are EA2. Applying the same argument to all points along γ_{bc}, we can "fill in" the triangle T:

PROPOSITION 35. *If X is locally compact and T is a triangle which satisfies* EA1 *and lies $B(p,r)$, where $B(p,3r)$ is a region of curvature bounded below, then T spans a surface in X which is isometric to the surface spanned by a representative $\widetilde{T}$ of T in S_k.*

The proofs of the next two simple but important lemmas are immediate from the way that angles are measured, together with A1, and Proposition 23, respectively.

LEMMA 36. *Let γ_i, α_i be minimal curves starting at $p_i \to p \in X$. If $\gamma_i \to \gamma$ and $\alpha_i \to \alpha$ for some minimal curves γ, α starting at p, then $\alpha(\gamma, \alpha) \leqslant \liminf \alpha(\gamma_i, \alpha_i)$.*

LEMMA 37. *If $\gamma_1, \gamma_2, \gamma_3$ are minimal curves starting at p, then $\alpha(\gamma_1, \gamma_2) + \alpha(\gamma_2, \gamma_3) + \alpha(\gamma_1, \gamma_3) \leqslant 2\pi$.*

COROLLARY 38. *If $\gamma_1, \gamma_2, \gamma_3$ are minimal curves starting at p, and $\pi - \alpha(\gamma_1, \gamma_3) \leqslant \varepsilon$ for some $\varepsilon \geqslant 0$, then $\alpha(\gamma_1, \gamma_2) + \alpha(\gamma_2, \gamma_3) \leqslant \pi + \varepsilon$.*

Although it is a trivial consequence of Lemma 37, we state the above corollary because we will use it in this form frequently; it allows us to bound an angle from *above* by bounding its (almost) complementary angle from below using A1. For example, this can be carried out at any point p which is in the interior of a geodesic, or is the endpoint of an almost extendable geodesic. By Theorem 27, this means almost everywhere. Note that the case $\varepsilon = 0$ is the "Theorem of Complementary Angles". A *branching* geodesic consists of two geodesics with common endpoint which coincide on an interval, but one is not an extension of the other. If such occurred in a space of curvature $\geqslant k$, then at the "branch point" p, the sum of angles between three geodesics leaving p would be $> 2\pi$. We obtain:

COROLLARY 39. *In X, geodesics do not branch (bifurcate). In particular, the restriction of any (minimal) γ_{ab} to a proper segment is the unique minimal curve joining its endpoints.*

COROLLARY 40. *Let Y be an inner metric space such that for every $p \in Y$, J_p contains a dense G_δ. If for some k, condition* A0 *holds for all triangles in any ball $B(x,r)$, then every quadruple in $B(x,r/2)$ isometrically embeds in S_m for some $m \geqslant k$.*

The last corollary, which implies the essential equivalence of Alexandrov's definition (together with suitably many minimal curves) and Berestovskii's definition, can be proved as follows: We can equivalently show that for any quadruple $\{a,b,c,d\}$ in $B(p,r/2)$, $V_k(a) \leqslant 2\pi$. By continuity and our assumption on J_p we can assume that all points can be joined by minimal curves, and now the conclusion follows from Lemma 37 and A1.

REMARK 41. We can now completely understand "one-dimensional" spaces X of curvature bounded below. By "one-dimensional", we mean that there is some point $p \in X$ that has at most two distinct unit geodesics leaving it. Since geodesics do not branch, there are exactly three possibilities: X is isometric to a closed interval, X is isometric to the reals, or X is isometric to a circle. Some of the arguments in the sequel require the existence of at least three minimal curves starting at every point. *From now on, we will assume, without further mention, that X is not one-dimensional.*

REMARK 42. It follows from the Rauch Comparison Theorem [23,27] that in a Riemannian manifold, sectional curvature $\geqslant k$ and A1 are equivalent; hence curvature $\geqslant k$ in the metric sense and sectional curvature $\geqslant k$ are equivalent.

3.4. *The global comparison theorem*

THEOREM 43. *If X is an inner metric space of curvature $\geqslant k$, then all of X is a region of curvature $\geqslant k$.*

The above theorem is known in Riemannian geometry as the Alexandrov–Toponogov Theorem, or simply Toponogov's Theorem. It was proved by Alexandrov for convex surfaces (using approximation by polyhedra) [2] and by Toponogov for Riemannian manifolds of any dimension [87]. The first completely metric proof of the theorem (even for Riemannian manifolds) may be found in [72], but the proof there requires geodesic completeness. In the above generality, proofs may be found in [76] and [19]. (Note that the proof in [18] uses local compactness; in the published version, the existence of minimal curves is assumed in a way that is not essential.) We outline the argument in [76], which in turn is a modified version of the proof in [72].

DEFINITION 44. For any $p \in X$, the *comparison radius at p* is defined to be $c_k(p) := \sup\{r\colon B(p,r) \text{ is a region of curvature} \geqslant k\}$.

If p does not lie in a region of curvature $\geqslant k$ then we can take $c_k(p) = 0$. The function c_k is obviously continuous, and proving the Global Comparison Theorem amounts to proving that $c_k(p) = \infty$ for some point p. We begin by describing a construction that goes back to Alexandrov, and is a standard step in proofs of Toponogov's Theorem in Riemannian geometry. In describing it, and in the proofs which follow, we adopt an American colloquialism and say that a triangle or hinge which satisfies the relevant curvature condition is AOK. Let p, q, r, s be points in X joined by minimal curves, where s is on the minimal curve joining q and r. We claim that if all the smaller hinges $H(\gamma_{qp}, \gamma_{qs})$, $H(\gamma_{sq}, \gamma_{sp})$, $H(\gamma_{sp}, \gamma_{sr})$ are AOK, then the large hinge $H(\gamma_{qp}, \gamma_{qr})$ is AOK. For the proof, choose a representative triangle $T(Q, P, S)$ of $T(q, p, s)$ in S_k and extend the shortest curve γ_{QS} beyond S to a geodesic Γ_{QR} of length $d(q,r)$. Then by A1, $\alpha(\gamma_{sq}, \gamma_{sp}) \geqslant \alpha_k(s; q, p) = \alpha(\gamma_{SQ}, \gamma_{SP})$, so $\alpha(\gamma_{sp}, \gamma_{sr}) \leqslant \alpha(\gamma_{SP}, \gamma_{SR})$ by Lemma 38, and it follows from A2 and the second monotonicity that $d(p,r) \leqslant d(P, R)$. On the other hand, A1 also implies that $\alpha(\gamma_{qp}, \gamma_{qr}) = \alpha(\gamma_{qp}, \gamma_{qs}) \geqslant \alpha(\gamma_{QP}, \gamma_{QR})$, so it follows from the second monotonicity that $H(\gamma_{qp}, \gamma_{qr})$ is AOK. This argument can be easily extended to an inductive argument in which the geodesic γ_{qr} has been subdivided finitely many times and each of the smallest resulting wedges except the "last" is AOK. In the proof, the process of always going back to the beginning angle after working one's way forward is more exaggerated. For that reason, we will refer to an argument of this type as a "b&f" (for backwards and forwards) argument. For example, a b&f argument can be used to prove the following proposition, by subdividing the nonminimal geodesic into minimal segments. Although it is of intrinsic interest, we will not use the proposition in this paper.

PROPOSITION 45. *Let $H(\gamma_{pq}, \gamma_{pr})$ be a hinge in a sufficiently large region of curvature $\geqslant k$ such that γ_{pq} is minimal and γ_{pr} is a (possibly nonminimal) geodesic of length $\leqslant \pi/\sqrt{k}$. Then $H(\gamma_{pq}, \gamma_{pr})$ is AOK.*

Since all points may not be joined by minimal curves, we need to use a modified form of the b&f argument. Suppose, in the above case, the point s is not joined to p by a minimal curve. Then by Corollary 30 we can find points $s_i \to s$ such that $s_i \in J_p \cap J_q \cap J_r$. Since $\alpha_k(s_i; q, r) \to \pi$, we can apply the same argument (with a few ε's introduced) to reach the same conclusion, as long as we know that the hinges $H(\gamma_{qp}, \gamma_{qs_i})$, $H(\gamma_{s_iq}, \gamma_{s_ip})$, $H(\gamma_{s_ip}, \gamma_{s_ir})$ are all AOK. We will still refer to this type of an argument as a b&f argument.

Theorem 43 follows from the next proposition. In fact suppose that $c_k(x) = r < \infty$ for some $r > 0$. Then Proposition 46 implies that there must exist a point $x_1 \in B(x, 3r)$ such that $c_k(x_1) < 1/6$. But then there must exist a point $x_2 \in B(x_1, 1/2)$ such that $c_k(x_2) < 1/12$. Continuing in this way we can construct a Cauchy sequence of points $\{x_i\}$ such that $c_k(x_i) \to 0$. But since X has curvature $\geqslant k$, $x := \lim x_i$ has $c_k(x) > 0$, a contradiction to the continuity of c_k.

PROPOSITION 46. *If there exists some $\kappa > 0$ such that $c_k \geqslant \kappa$ on the ball $B(p, r)$, then $c_k(p) \geqslant r/3$.*

The basic idea of the proof is to fix a base point $q \in B(p, r/3)$. Choose positive $\chi < \kappa/8$ such that if (γ_{ab}, τ_{ac}) is a hinge in S_k such that $\alpha(\gamma, \tau) \leqslant \pi/2$ and $L(\tau) < 4\chi$ then for all $t \in [0, L(\tau)]$, $d(a, \tau(t)) \leqslant L(\gamma) + 2\chi$. The last condition can be accomplished because, as was pointed out after formula 4, small and thin triangles in S_k are almost Euclidean. Of course the choice of χ depends on, and is an increasing function of, k. Then for some suitably chosen small number $\chi > 0$ we want to prove the following statement by induction:

S(n) If $(a; q, b)$ is a triple such that $q, b \in J_a$, $d(q, b), d(q, a) < n\chi$ and $d(a, b) < \chi$, the hinge $(\gamma_{aq}, \gamma_{ab})$ is AOK.

If we can prove this statement, then we can remove the requirement that $d(a, b) < \chi$ by a b&f argument. One only has to be careful in the case of $k > 0$, as the induction takes one close to $n\chi = \pi/\sqrt{k}$, because then the existence of representatives in S_k becomes a problem. By making χ small, we can come arbitrarily close to $\pi/\sqrt{k}$. Now suppose that $d(a, p) = \pi/\sqrt{k}$ and a and p are joined by a minimal curve γ. If $q \in J_a \cap J_p \cap J_{p_i}$, where $p_i \to p$ along γ, then applying S(n) for small χ and passing to the limit shows that the minimal curves γ_{aq} and γ_{pq} together form a minimal curve from a to p. Since our space is not one-dimensional (see Remark 41), we could choose q not on γ, so there are two distinct minimal curves joining a and p. Since geodesics do not bifurcate and a and p were arbitrary of distance $\pi/\sqrt{k}$, we see that the diameter of X is at most $\pi/\sqrt{k}$. It is now possible to argue that any two geodesics starting at p must come back together at a, and complete the proof.

S(n) is proved as follows. The case $n = 1$ is trivial, since χ is small. Suppose S(n) holds, and we have a triple $(a; q, b)$ such that $q, b \in J_a$, $d(q, b), d(q, a) < (n+1)\chi$ and $d(a, b) < \chi$. We go back along γ_{qa} to a point c such that $d(a, c) = 3\chi$. Now χ was chosen small enough that the new triple $(a; c, b)$ lies in a ball of radius 2ε, so we will be done by a b&f argument if we can show that any triple $(d; q, b)$ such that

$$q, b \in J_d, \quad d(q, d) < (n-2)\chi \quad \text{and} \quad d(a, b) < 4\chi \tag{5}$$

is AOK.

Let $\tau = \gamma_{qd}$ and $\eta = \gamma_{db}$ and suppose (T, N) represents the hinge (τ, η) in S_k. We denote points in S_k by the capitals of the letters denoting points in X. Let $\theta = \gamma_{BQ}$ in S_k. Note that if $d(Q, B) \geqslant n\chi > d(q, b)$ then the hinge is A2 and we are finished, so we suppose that $d(Q, B) < n\chi$. Now we can choose χ small enough that the hinge (T, N) is sufficiently thin (recall that thin triangles in S_k are approximately Euclidean, or see [76] for more details) that the point F on θ such that $d(B, F) = 2\chi$ has the following property: If Ω is minimal from D to G then for all t, $d(Q, \Omega(t)) < n\chi$.

To motivate the rest of the proof, we first show how, in a Riemannian manifold, the proof is now easily finished. Choose a geodesic ω starting at d such that the direction of ω is between the directions of η and τ, and $\alpha(\omega, \tau) = \alpha(\Omega, T)$. Let $f = \omega(d(D, F))$. Then a b&f argument shows inductively that ω stays inside $B(p, n\chi)$, that the hinge (ω, τ) is AOK. The hinge (ω, η) is small enough to be AOK, and now we get that (τ, η) is AOK from the fact that $\alpha(\eta, \tau) = \alpha(\omega, \tau) + \alpha(\omega, \eta)$ and $d(b, q) \leqslant d(b, f) + d(f, q)$.

In the general case, we will see later (Proposition 53) that there is a *direction* ω having the above properties, but it may not be a geodesic direction – and even if it were, it might not extend to sufficient length. Instead, we show the existence of points which play the role of points on ω in the above argument. This kind of idea is carried further by the construction of gradient curves and quasi-geodesics later in this paper. The actual argument is somewhat technical; for our sketch here we assume that all points can be joined by minimal curves, and limit ourselves to showing the following: There exist points $d' \in X$ and $D' \in S_k$ such that

(1) $\alpha(d'; b, q) \leqslant \alpha(D'; B, Q)$,
(2) $d(q, d') = d(Q, D') = d(q, d)$,
(3) $d(d', b) \leqslant d(D', B)$,
(4) $\varepsilon(D'; B, Q) < \varepsilon(D; B, Q)$.

This is the first, and essentially the "open" step in an open-closed argument that shows we can find points d'' and D'' satisfying conditions (2), (3), and $\varepsilon(D'; B, Q) = 0$. We then have

$$d(b,q) \leqslant d(b,d'') + d(d'',q) \leqslant d(B,D'') + d(D'',Q) = d(B,Q).$$

(The condition (1) is needed to ensure we can iterate the construction.) To construct D' and d', choose a point E on γ_{QD} such that $d(E,D) = \chi$ and let D' be the point on γ_{EB} such that $d(D',Q) = d(d,q)$. Now construct d' in the same way in S_k. By construction, (2) and (4) are satisfied. In what follows it is necessary to use triangle comparisons; in each case, the comparisons are valid because the triangles or hinges involved are either small enough or satisfy the inductive hypothesis. From A2 we get that $d(e,b) \leqslant d(E,B)$. From this inequality and A1 we get $\alpha(e;d,b) \geqslant \alpha(E;D,B)$, and therefore $\alpha(e;q,d') \leqslant \alpha(E;Q,D')$. From A1 we now obtain that $d(e,d') \geqslant d(E,D')$, which implies (3). From $d(e,d') \geqslant d(E,D')$ and A1 we get $\alpha(d'e) \geqslant \alpha(D';e,Q)$, which implies (1).

COROLLARY 47. *If X is a space of curvature $\geqslant k > 0$ then* $\operatorname{diam}(X) \leqslant \pi/\sqrt{k}$. *If in addition X is locally compact, then X is compact.*

3.5. *The space of directions*

In this section we assume that X is an inner metric space of curvature $\geqslant k$. We now begin the construction of the infinitesimal structure of a space of curvature bounded below. At any point p we consider the space S_p of all unit geodesics starting at p. We identify geodesics which agree on some initial interval; equivalently, for our space we can take all unit parameterized geodesics starting at p with maximal domain of definition. It is easy to verify that the angle α between geodesics is a metric on S_p; S_p with this metric is called the *space of geodesic directions*. We denote by Σ_p the metric completion of S_p, which is called the *space of directions* at p. Note that in a Riemannian manifold, M^n, $\Sigma_p = S_p = S^{n-1}$, and is identified with the unit sphere in the tangent space.

REMARK 48. We will abuse notation, writing $\gamma \in S_p$ to mean simply that γ is some unit parameterized minimal curve starting at p, such that $\gamma(0) = p$. From now on, when considering a minimal curve γ_{ab}, we will represent, without further comment, the geodesic direction corresponding to γ_{ab} (in either S_a or S_b) by γ.

PROPOSITION 49. *For any $p \in X$, if Σ_p is an inner metric space then all of Σ_p is a region of curvature $\geqslant 1$.*

The essential argument for the above proposition may be found in Proposition 2.4 [72] (the additional assumptions in [72] can be removed by using Corollary 38 in this chapter, rather than Lemma 2.3 in [72]). We prove the statement here under the simplifying assumption that every pair of points in S_p can be joined by a minimal curve; such curves can often be found (cf. Proposition 53). Let $\gamma_1, \ldots, \gamma_4 \in S_p$ be such that $\gamma_1, \gamma_2, \gamma_3$

lie on a minimal curve in S_p. We also assume $\alpha_{ij} := \alpha(\gamma_i, \gamma_j) < \pi$. Fix a base point P in S_k^3 and choose geodesics $\Gamma_1, \ldots, \Gamma_4$ starting from P such that $\alpha(\Gamma_i, \Gamma_j) = \alpha_{ij}$ for all $i \leqslant j$ except $i = 3$, $j = 4$. Since the directions Γ_i are isometrically embedded in the space of directions Σ_P, which is isometric to S_1, by monotonicity we can verify A2 by proving that $\alpha(\Gamma_3, \Gamma_4) \geqslant \alpha_{34}$. Let $t_i \to 0$ and let $a_{ij} := \gamma_j(t_i)$ and $A_{ij} := \Gamma_j(t_i)$. Then $d(a_{ij}, a_{ik}) = d(A_{ij}, A_{ik}) + o(t_i)$. Let ϕ^i_{jk} be minimal in X from a_{ij} to a_{ik} and Φ^i_{jk} be minimal in S_k^3 from A_{ij} to A_{ik}; then we have by A2 that $\alpha(\phi^i_{21}, \phi^i_{24}) \geqslant \alpha(\Phi^i_{21}, \Phi^i_{24}) + o(t_i)$, and so $\alpha(\phi^i_{23}, \phi^i_{24}) \leqslant \alpha(\Phi^i_{23}, \Phi^i_{24}) + o(t_i)$. It now follows from A1 that $d(A_{i3}, A_{i4}) \geqslant d(a_{i3}, a_{i4}) + o(t_i)$, and therefore that $\alpha(\Gamma_3, \Gamma_4) \geqslant \alpha_{34}$.

In many important cases, which we will consider later, Σ_p is known to be an inner metric space. As this article was going to press we received a copy of the thesis of Stephanie Gloor (1998, Zurich), which contains an example of an inner metric space with curvature $\geqslant k$ such that the space of directions at some point is not an inner metric space. This example leaves open the question of whether the space of directions must still have curvature $\geqslant 1$ in the sense of Berestovskii–Wald, and whether (as is probably the case) the metric still has some weaker geometric property related to being an inner metric.

The above proposition shows that the infinitesimal properties of spaces of curvature $\geqslant k$ are intimately connected with the global properties of spaces of curvature $\geqslant 1$. We need a few more basic results.

LEMMA 50. *Suppose $p, q \in X$ and $p \in J_q$. Then for any $q_i \to q$ and minimal curves γ_{pq_i}, $\alpha(\gamma_{pq_i}, \gamma_{pq}) \to 0$.*

To prove the lemma, note that by Lemma 38, Lemma 15 and A1 it suffices to observe that if p' is such that $\sigma(p; q, p')$ is close to π then for i large, $\sigma(p; q_i, p')$ is close to π. The next proposition is a generalization of Satz 8, Section 38 in [78].

PROPOSITION 51. *Suppose $p, q \in X$ and $p \in J_q$. For any minimal curve γ starting at q, if β is the (unique) unit minimal curve from q to p, then $\alpha(\gamma, \beta) = \lim_{t \to 0} \alpha_k(q; \gamma(t), p)$.*

PROOF. If β and any segment of γ together form a minimal curve, or if $\alpha(\beta, \gamma) = 0$ then the proof is immediate; we assume the contrary. It is immediate from A1 that $\alpha(\gamma, \beta) \geqslant \lim_{t \to 0} \alpha_k(q; \gamma(t), p)$; we need to show the opposite inequality. Fix $\varepsilon > 0$ and choose (using the definition of the angle) $a := \gamma(s)$ and $b := \beta(s) \neq p$ such that $\alpha_k(q; a, b) > \alpha(\gamma, \beta) - \varepsilon$. Let $D := d(p, b) > 0$ and fix positive $t < s$ small enough that $d(p, \gamma(t)) > D$.

We make the following claim: We can find a point $u \in J_p \cap J_a \cap J_q$ arbitrarily close to $\gamma(t)$ with the following properties:

(1) For the point c on γ_{pu} such that $d(c, p) = D$, $\alpha_k(q; u, c) \leqslant \alpha_k(q; \gamma(t), p)$.

(2) $\alpha_k(q; u, c) > \alpha_k(q; a, c) - \varepsilon$.

To prove the claim, suppose we take a sequence $u_j \in J_p \cap J_a \cap J_q$ converging to $\gamma(t)$ and the points c_j on $\gamma_j := \gamma_{pu_j}$ such that $d(p, c_j) = D$. First, there exists an $\omega > 0$ such that for all large j, $\alpha_k(q; c_j, p) \geqslant \omega$. Suppose, on the contrary, $\alpha_k(q; c_j, p) \to 0$. Then $\alpha_k(c_j; p, q) \to \pi$ and $\alpha(\gamma_{c_j p}, \gamma_{c_j q}) \to \pi$. Then $\alpha(\gamma_{c_j u_j}, \gamma_{c_j q}) \to 0$. If $\limsup d(c_j, u_j) \geqslant d(b, q)$ then (taking a subsequence if necessary) for some q_j on $\gamma_{c_j u_j}$, $q_j \to q$. By Lemma 50, $\alpha(\gamma_{pu_j}, \beta) \to 0$ and, since $\gamma(t) \neq q, \alpha_k(q; p, \gamma(t)) \to \pi$, a contradiction to $\alpha(\beta,\gamma)\neq\pi$. A similar argument contradicts $\alpha(\beta,\gamma)\neq 0$ when $\limsup d(c_j,u_j)\geqslant d(b,q)$. We now choose j large enough that $|\alpha_k(q;u_j,p)-\alpha_k(q;\gamma(t),p)|\leqslant\omega$ and u_j is close to $\gamma(t)$. Take representative triangles $\widetilde{T}(\tilde{q},\tilde{u}_j,\tilde{c}_j)$ and $\widetilde{T}(\tilde{q},\tilde{c}_j,\tilde{p})$ having a common side between points $\tilde{q}$ and $\tilde{c}_j$. By the triangle inequality, $d(\tilde{u}_j,\tilde{p})\leqslant d(\tilde{u}_j,\tilde{c}_j)+d(\tilde{c}_j,\tilde{p})=d(u_j,p)$. In other words, to make a representative triangle of $T(u_j,p,q)$, we need to rotate the segment between $\tilde{q}$ and $\tilde{p}$ to make the angle at $\tilde{q}$ larger (or at least not smaller). Now we have that

$$\begin{aligned}\alpha_k\big(q;\gamma(t),p\big)&\geqslant\alpha_k(q;u_j,p)-\omega\geqslant\alpha_k(q;u_j,c_j)+\alpha_k(q;c_j,p)-\omega\\&\geqslant\alpha_k(q;u_j,c_j)\end{aligned}$$

which proves part (1).

To prove part (2), observe that A0 implies that $\alpha_k(q;\gamma(t),c)\geqslant\alpha_k(q;a,c)$, so we can satisfy (2) simply by taking u_j close enough to $\gamma(t)$.

To finish the proof of the proposition, choose $t_i\to 0$, u_i and c_i satisfying (1) and (2) above. By choosing u_i close enough to $\gamma(t_i)$ we can ensure that $u_i\to q$; Lemma 50 and A2 now imply that $c_i\to b$, and $\alpha_k(q;a,c_i)\to\alpha_k(q;a,b)$, so for large enough i, $\alpha_k(q;a,c_i)\geqslant\alpha(\gamma,\beta)-2\varepsilon$. Putting our inequalities together we get

$$\alpha_k\big(q;\gamma(t_i),p\big)\geqslant\alpha_k(q;u_i,c_i)\geqslant\alpha_k(q;a,c_i)-\varepsilon\geqslant\alpha(\gamma,\beta)-3\varepsilon.$$

Since the sequence $\{t_i\}$ and ε were arbitrary, the proof is complete. □

The only place in the above proof where we used the fact that $p\in J_q$ was when we needed to know that $\alpha(\beta,\gamma_{pu_j})\to 0$. Therefore the same proof (even a simpler one, since we don't have to pick the points u_j) gives us the next proposition. We will need it somewhat later.

PROPOSITION 52. *Let $p\neq q$ in X, and let β join p and q. Suppose that γ is a minimal curve starting at q and for some $t\to 0$, there exist minimal curves β_t from p to $\gamma(t)$ such that $\alpha(\beta,\beta_t)\to 0$. Then $\alpha(\gamma,\beta)=\lim_{t\to 0}\alpha_k(q;\gamma(t),p)$.*

PROPOSITION 53. *Let $p,a,b\in X$ and suppose that $p\in J_a\cap J_b$. Then there exists a minimal curve in Σ_p joining γ_{pa} and γ_{pb}.*

For each $t>0$ we join $\gamma(t)$ and $\beta(t)$ by a minimal curve ξ_t. Let β_t and γ_t denote the restrictions of the two curves to $[0,t]$. Our first claim is that the sum of the angles $\sum(t)$ of the triangle formed by γ_t,β_t,ξ_t tends to π as $t\to 0$; that is, the triangle becomes Euclidean in the limit. To see this, represent the hinge (γ,β) by $(\tilde{\gamma},\tilde{\beta})$ in S_k, and let $\tilde{\xi}_t$ denote the minimal curve joining $\tilde{\gamma}(t)$ and $\tilde{\beta}(t)$. We give a sketch of the argument, using some simplifying assumptions: we assume that every pair of points can be joined by a minimal curve, and that $\gamma:=\gamma_{pa}$ and $\beta:=\gamma_{pb}$ are in fact extendable beyond p. In the proof of Proposition 51 we showed the kind of unpleasant extra details needed if we do not use these simplifying assumptions. Also, we will limit our argument to constructing a midpoint between β and γ. Note that, by moving a and b along

γ and β, respectively, we can assume $a, b \in J_p$. Then $L(\tilde{\xi}_t) - L(\xi_t) = o(t)$. From Proposition 17, we see that $\liminf_{t\to 0} \sum(t) \geqslant \pi$. On the other hand, Proposition 51 implies that $|d(a, \beta(t)) - d(\tilde{a}, \tilde{\beta}(t))| = o(t)$. From Proposition 17 and A1 we see that if $\alpha'(t)$ denotes the angle of the hinge complementary to (γ_t, ξ_t) and $\tilde{\alpha}'(t)$ denotes the angle of the hinge complementary to $(\tilde{\gamma}_t, \tilde{\xi}_t)$ then $\liminf_{t\to 0} \alpha'(t) \geqslant \lim_{t\to 0} \tilde{\alpha}'(t)$. We immediately get that $\limsup_{t\to 0} \alpha(\gamma_t, \xi_t) \leqslant \lim_{t\to 0} \alpha(\tilde{\gamma}_t, \tilde{\xi}_t)$. Now $\liminf_{t\to 0} \alpha(\gamma_t, \xi_t) \geqslant \lim_{t\to 0} \alpha(\tilde{\gamma}_t, \tilde{\xi}_t)$ follows from Proposition 17 and A1. The same argument applies to $\lim_{t\to 0} \alpha(\beta_t, \xi_t)$, and we have proved the claim.

Let and let $m_t, \tilde{m}_t$ denote the midpoint of $\xi_t, \tilde{\xi}_t$ and let $\eta_t, \tilde{\eta}_t$ be a minimal curve from $p, \tilde{p}$ to $m_t, \tilde{m}_t$, respectively. We complete the proof by showing $\lim_{t\to 0} \alpha(\eta_t, \gamma) = \alpha(\gamma, \beta)/2$ (a similar argument shows that $\lim_{t\to 0} \alpha(\eta_t, \beta) = \alpha(\gamma, \beta)/2$). By A1 and Proposition 17 we get $\liminf_{t\to 0} \alpha(\eta_t, \gamma) \geqslant \alpha(\gamma, \beta)/2$. To get the opposite inequality, recall that we have assumed that γ is extendable (as a minimal curve) beyond p. Let $c \in J_p$ be a point on that extension, and let ζ_t be minimal from c to m_t, with corresponding point $\tilde{c}$ and curve $\tilde{\zeta}_t$ in S_k. Now, since $\lim_{t\to 0} \alpha(\gamma_t, \xi_t) = \lim_{t\to 0} \alpha(\tilde{\gamma}_t, \tilde{\xi}_t)$, we get from A2 and Proposition 17 that $L(\zeta_t) - L(\tilde{\zeta}_t) \leqslant o(t)$. That $|L(\zeta_t) - L(\tilde{\zeta}_t)| \leqslant o(t)$ can be obtained by observing that by Proposition 51 $|d(c, \beta(t)) - d(\tilde{c}, \tilde{\beta}(t))| = o(t)$ and applying A0 the triangle with corners $c, \gamma(t), \beta(t)$. Proposition 17 and A1 now imply that $\lim_{t\to 0} \alpha(\eta_t, \gamma_{pc}) \geqslant \pi - \alpha(\tilde{\gamma}, \tilde{\beta})/2 = \pi - \alpha(\gamma, \beta)/2$, and the proof is complete by Corollary 38.

REMARK 54. When we refer to the unit sphere S^n (or simply an n-sphere) we mean the sphere of curvature 1 when $n \geqslant 2$, a circle of length 2π when $n = 1$, and two points of distance π when $n = 0$.

If γ_{ab} is almost extendable past b, by Theorem 27 we can find a point c joined to b by a minimal curve, such that $\sigma(b; a, c)$ is arbitrarily small. But then $\alpha_k(b; a, c)$ is close to π, so by A1, $\alpha(\gamma_{ba}, \gamma_{bc})$ is arbitrarily close to π. Given another such point c', we have from Lemma 38 that $\alpha(\gamma_{bc'}, \gamma_{bc})$ is small. In this way we can construct a Cauchy sequence of directions γ_i such that $\alpha(\gamma, \gamma_i) \to \pi$. We have proved:

PROPOSITION 55. *If γ_{ab} is almost extendable past b, then γ has a complement in Σ_p; i.e., a direction $-\gamma \in \Sigma_b$ such that $\alpha(\gamma, -\gamma) = \pi$.*

Suppose now that $\gamma \in S_p$ has a complement. First we restrict γ to a segment γ_{pq} which is extendable past q. We then choose geodesic directions $\gamma_i \to \gamma$. Now by choosing points x_i sufficiently close to p along γ_i, we can ensure that $\alpha_k(p; q, x_i) \to \pi$. In this way we can prove a converse to Proposition 55:

PROPOSITION 56. *If $\gamma \in S_p$ has a complement $-\gamma$ in S_p, then there is a restriction of γ that is almost extendable past p (and is the unique minimal curve between its endpoints).*

Ordinarily, the angle between minimal curves is only semicontinuous (cf. Lemma 36). Suppose that γ_i, α_i start at p_i and $\gamma_i \to \gamma$, $\alpha_i \to \alpha$, where γ, α are minimal starting at

$p = \lim p_i$. Suppose that α is extendable past p, and denote by $-\alpha_{pq}$ a minimal curve starting at p, $q \in J_p$, such that $\alpha(\alpha, -\alpha) = \pi$. Let β_i be minimal from p_i to q. By Lemma 36, $\alpha(\alpha_i, \beta_i) = \pi$. Then by Lemma 38, $\alpha(\gamma, \alpha) = \pi - \alpha(\gamma, -\alpha) \geqslant \pi - \liminf \alpha(\gamma_i, \beta_i) = \limsup \alpha(\gamma_i, \alpha_i)$. Since we already know the opposite inequality, we have proved the continuity of the angle. With a little more work one can show:

PROPOSITION 57. *Let $\gamma_i, \alpha_i \in S_{p_i}$, where $p \in X_i$, and $p_i \to p \in X$. If $\gamma_i \to \gamma \in S_p$, and $\alpha_i \to \alpha \in S_p$ and α is almost extendable past p then $\alpha(\gamma, \alpha) = \lim \alpha(\gamma_i, \alpha_i)$.*

3.6. *Differentiability of the distance function*

DEFINITION 58. Let X be a space of curvature $\geqslant k$ and $p \in X$. Suppose $\gamma : [a, b] \to X$ is unit minimal. We define $d_p : [a, b] \to X$ by $d_p(t) = d(p, \gamma(t))$.

The differentiability properties of the distance function $d_p(t)$ in a space of curvature $\geqslant k$ are of fundamental importance. In a Riemannian manifold, if there is a unique minimal curve β from p to $\gamma(t)$, $t \in (a, b)$ then $d'_p(t) = -\cos\alpha(\gamma|_{[t,b]}, \beta)$ (in S_k this follows easily from the cosine laws). The second derivative of $d_p(t)$ is (unlike the first) dependent on curvature and the distance from p to $\gamma(0)$. We will consider analytical properties of the distance function further in the next section. For the moment we limit ourselves to the following results.

DEFINITION 59. Let $x, y, z \in X$. For every $\varepsilon > 0$, denote by $\alpha_\varepsilon(y; x, z)$ the infimum of $\alpha(\gamma_{yx'}, \gamma_{yz'})$, where the infimum is taken over all x', z' such that $d(x, x'), d(z, z') < \varepsilon$. Let $\alpha(y; x, z) = \lim_{\varepsilon \to 0} \alpha_\varepsilon(y; x, z)$.

The proof of the following lemma is now immediate. It allows us to pick the points x' and z' in J_q, and we will use it without further reference.

LEMMA 60. *Let $x, y, z \in X$. For every $\varepsilon > 0$, denote by $\alpha^J_\varepsilon(y; x, z)$ the infimum of $\alpha(\gamma_{yx'}, \gamma_{yz'})$, where the infimum is taken over all x', z' such that $d(x, x'), d(z, z') < \varepsilon$ and $x', z' \in J_q$. Then $\alpha(y; x, z) = \lim_{\varepsilon \to 0} \alpha^J_\varepsilon(y; x, z)$.*

PROPOSITION 61. *Let $p \neq q \in X$ and suppose that $\gamma \in S_q$ and $r \neq p$ is in the interior of γ. Then $d_p(t)$ is differentiable from the right at $t = 0$, and $d'_p(0) = -\cos\alpha(p, \gamma)$, where $\alpha(p, \gamma) = \lim_{t \to 0} \alpha_k(q; p, \gamma(t)) \leqslant \alpha(q; p, r)$.*

PROOF. First note that the function $f(t) = \alpha_k(q; p, \gamma(t))$ is bounded and decreasing by A0 and the second monotonicity, so the above limit exists. In S_k, consider a minimal curve Γ_{PQ} such that $L(\Gamma) = d(p, q)$. Now rotate Γ, say, clockwise, at Q to a minimal curve Γ_t such that $\alpha(\Gamma, \Gamma_t) = \alpha_k(q; p, \gamma(t))$. Note that $\Gamma_0 := \lim_{t \to 0} \Gamma_t$ satisfies $\alpha(\Gamma_0, \Gamma) = \alpha(p, \gamma)$ and $d(P, \Gamma_t(t)) = d_p(t)$. Then since $\alpha(\Gamma_t, \Gamma_0) \to 0$,

$$\left|d\big(P, \Gamma_t(t)\big) - d\big(P, \Gamma_0(t)\big)\right| \leqslant d\big(\Gamma_t(t), \Gamma_0(t)\big) = o(t).$$

Now

$$\lim_{t\to 0}\frac{d(P,\Gamma_0(t))-d(P,Q)}{t}=-\cos\alpha(p,\gamma);$$

we get that $d'_p(0)=-\cos\alpha(p,\gamma)$. The final inequality follows from the definition of $\alpha(q;p,r)$ and A1. □

From Proposition 51 we get the following:

COROLLARY 62. *Let $p\neq q\in X$ with $p\in J_q$. For any minimal γ starting at q, the right derivative $d'_p(0)=-\cos\alpha(\gamma,\beta)$, where β is the minimal curve joining p and q.*

The following result is useful for computing angles when the points in question are not moving along some fixed minimal curve. Its proof can be found in [74].

PROPOSITION 63. *Let $p\in X$ and suppose that γ_{px} is the unique minimal curve between its endpoints. Suppose that $a_i,b_i\to p$, γ_i is minimal from a_i to x and β_i is minimal from a_i to b_i. Then $\alpha=\lim_{i\to\infty}\alpha(\gamma_i,\beta_i)$ exists if and only if*

$$L=\lim_{i\to\infty}\frac{d(x,b_i)-d(x,a_i)}{d(a_ib_i)}$$

exists. If L and α exist, then $L=-\cos\alpha$.

From Proposition 61 and Lemma 38, we get the following:

LEMMA 64. *Suppose $p\in X$ and γ: $[-t,t]$ is a geodesic with $p\neq\gamma(0)$. Then if $\gamma(0)$ is a local minimum for $d_p(t):=d(p,\gamma(t))$ then for any minimal β from $\gamma(0)$ to p, $\alpha(\gamma,\beta)=\pi/2$.*

4. Constructions

The importance of a lower curvature bound is partly due to the fact that it is preserved under a great variety of constructions, many of which, such as cones and joins, generally fail to preserve the property of being a Riemannian manifold – or even a topological manifold. The basic constructions given here were noticed by various people, sometimes independent of one another. Gluings, and scaling, and products are "classical" constructions; cone constructions were first discovered by Berestovskii, who used them in his solution of Borsuk's problem of metrizing polyhedra [9]. We remind the reader of our exclusion of one-dimensional spaces (cf. Remark 41), which often require separate but trivial treatment. For example, Propositions 87 and 90 are only true as stated if the cone is not 1-dimensional. If the cone is 1-dimensional then X is a one or two point space.

4.1. *Induced metrics, gluing, scaling*

DEFINITION 65. X is called *pre-inner* if every pair of points in X can be joined by a rectifiable curve, and for every $p \in X$ and $\delta > 0$ there exists an $\varepsilon > 0$ such that if $y \in B(p, \varepsilon)$ then y is joined to p by a curve of length less than δ.

A pre-inner metric space (X, d) has an *induced inner metric* d_I, where $d_I(x, y) = \inf\{L(c)\colon c \text{ joins } x \text{ and } y\}$. It follows from Definition 65 that d_I (which always satisfies $d_I \geqslant d$) has the same topology as d. A special case is when Y is a subset of an inner metric space X so that the subspace metric on Y is pre-inner; we refer to the inner metric induced by the subspace metric on Y as simply the *induced inner metric* d_I on Y. We do not know of any purely topological conditions that are necessary and sufficient for the existence of an inner metric on a topological space.

DEFINITION 66. If Y is a subset of X, then Y is called *metrically embedded* if the restriction of the metric of X to Y is an inner metric on Y.

Let (X_1, d_1) and (X_2, d_2) be metric spaces and $Y_1 \subset X_1$, $Y_2 \subset X_2$ be closed subsets such that (with the subspace metric) there is an isometry $I : Y_1 \to Y_2$. We "glue" X_1 to X_2 using I in the following way. First, we take the disjoint union $X_1 \amalg X_2$ and define a space X by identifying the points in $y_1 \in Y_1$ and $y_2 \in Y_2$ such that $I(y_1) = y_2$. We have a quotient map $q : X_1 \amalg X_2 \to X$; we denote by Y the set $q(Y_1) = q(Y_2)$, and identify X_i with $q(X_i)$. Then the space $X_1 \cup_I X_2$ is obtained by putting on X the following metric: $d(a, b) = d_i(a, b)$ if a and b both lie X_i for fixed $i = 1, 2$, or $d(a, b) = \inf_{y \in Y}\{d_1(a, y) + d_2(y, b)\}$ if $a \in X_1$ and $b \in X_2$. It is not hard to verify d is a well-defined metric, and the subspace metric on X_i is the same as d_i. The following proposition is not hard.

PROPOSITION 67. *Let X_1, X_2 be inner metric spaces, $Y_1 \subset X_1$, $Y_2 \subset X_2$ be closed subsets, and $I : Y_1 \to Y_2$ be an isometry. Then $X_1 \cup_I X_2$ is an inner metric space, and X_i is metrically embedded in $X_1 \cup_I X_2$.*

Generally, gluing sets with a lower curvature bound does not result in a space with curvature bounded below, even when the sets Y_i are metrically embedded. For example, one can glue two triangles along a side of equal length, identifying two vertices with angles $> \pi/2$. Then the glued edges form a minimal curve that branches along the two remaining edges at each endpoint, so the glued space cannot have a lower curvature bound.

If (Y, d) is a metric space, we can *scale* Y by the positive factor c to produce the new metric space (Y, cd). Clearly this operation preserves the property of being an inner metric space. Scaling a space form S^n_m by a factor of c results in the space form S^n_{m/c^2}, and so it follows from Definition 22 that scaling a space of curvature $\geqslant k$ results in a space of curvature $\geqslant k/c^2$.

4.2. *Gromov–Hausdorff convergence*

Recall that if A, B are subsets of a metric space Z, then the *Hausdorff distance* between A and B is defined to be

$$d_H(A, B) = \inf\{\varepsilon > 0\colon\ A \subset B_\varepsilon(B) \text{ and } B \subset B_\varepsilon(A)\},$$

where $B_\varepsilon(A) = \{x \in Z\colon\ d(x, A) < \varepsilon\}$.

DEFINITION 68. Let W, Y be compact metric spaces. The *Gromov–Hausdorff distance* between W and Y is defined to be

$$d_{GH}(W, Y) = \inf\{d_H(\phi_W(W), \phi_Y(Y))\},$$

where the infimum is taken over all metric spaces Z with isometric embeddings $\phi_W : W \to Z$ and $\phi_Y : Y \to Z$.

The Gromov–Hausdorff metric is a *bona fide* metric on the space of all compact metric spaces, if isometric spaces are considered equivalent. In general, the Gromov–Hausdorff metric is very weak, and says little about the structure of close spaces. However, in classes of metric spaces with sufficient bounds on geometric quantities, the Gromov–Hausdorff metric achieves two important properties: it is totally bounded (i.e., has compact metric completion, or is "precompact"), and spaces close in the Gromov–Hausdorff metric are topologically "close". We will discuss these properties later in this article.

We will be interested in convergent sequence of metric spaces in the Gromov–Hausdorff metric, and will now establish our notation. From now on, $X_i \to X_0$ will denote that the metric spaces X_i converge to X_0 in the Gromov–Hausdorff metric. By the results of [32] we can assume that the spaces X_i are all subspaces of a fixed metric space Z, and convergence is in the original Hausdorff sense. It is easy to show that for each $p, q \in X_0$ there exist points $p_i, q_i \in X_i$ such that $p_i \to p$, $q_i \to q$, and given any such sequences $p_i \to p$ and $q_i \to q$, $d_i(p_i, q_i) \to d_0(p, q)$. From these facts we immediately have the following lemma:

LEMMA 69. *Let* $X_i \to X_0$ *(*X_i *compact metric spaces).*
(1) *If each* X_i, $i > 0$, *is an inner metric space then* X_0 *is an inner metric space.*
(2) *If* $B(p_i, r) \subset X_i$ *is a region of curvature* $\geqslant k$ *for all* $i > 0$ *and* $p_i \to p \in X_0$, *then* $B(p, r)$ *is a region of curvature* $\geqslant k$ *in* X_0.

REMARK 70. It follows from Theorem 43 and the above lemma that X_0 has curvature $\geqslant k$ if each X_i has curvature $\geqslant k$. In other words, we see that the class of all compact inner metric spaces of curvature $\geqslant k$ and $\operatorname{diam}(X) \leqslant D$ (for fixed k and D) is closed under the Gromov–Hausdorff metric.

Given curves $\psi_i : [a, b] \to X_i \subset Z$, we will write $\psi_i \to \psi_0$ if ψ_i converges uniformly to ψ_0 in (the ambient metric space) Z. Clearly, if each ψ_i is minimal, then so is ψ_0. More generally, it follows from the way in which lengths are measured that $\liminf L(\psi_i) \geqslant L(\psi_0)$. It is not always possible to approximate minimal curves in X_0 by minimal curves in X_i. A simple counter example can be constructed as follows: Let S be a circle and a_1 and a_2 be antipodal points. Fix a semicircle Σ determined by a_1 and a_2, and points $a_{ij} \to a_j$ $(j = 1, 2)$ in Σ. Form inner metric spaces S_i by attaching segments σ_1, σ_2 of fixed length to a_{i1}, a_{i2}, respectively. In the limit we attach σ_i to a_i, and there are two minimal curves from the midpoint of σ_1 to the midpoint of σ_2, one passing through Σ and one not. The latter minimal curve cannot be approximated by minimal curves in the sequence S_i. However, if X_0 has curvature $\geqslant k$, then by Corollary 39, each minimal curve in X_0 has a restriction that is the unique minimal curve γ between its endpoints, say p and q. But then choosing $p_i \to p$ and $q_i \to q$ and $\gamma_i = \gamma_{p_i q_i}$, it follows from uniqueness that $\gamma_i \to \gamma$. With a little more care (cf. [74]) we obtain the following:

LEMMA 71. *Let $X_i \to X_0$, and suppose X_0 has curvature $\geqslant k$. Then for each γ_{ab}, γ_{ac} in X_0 there exist $a_i \to a$ and $\gamma_{a_i b_i}, \gamma_{a_i c_i}$ in X_i such that $\gamma_{a_i b_i} \to \gamma_{ab}$ and $\gamma_{a_i c_i} \to \gamma_{ac}$.*

DEFINITION 72. The *dilatation* of a mapping $f : A \to B$ between metric spaces A and B is defined to be

$$\operatorname{dil}(f) := \log\left(\sup_{x \neq y \in A} \frac{d(f(x), f(y))}{d(x, y)} \right).$$

Note that a mapping is distance nonincreasing if and only if it has nonpositive dilatation.

DEFINITION 73. If A and B are homeomorphic metric spaces, the *Lipschitz* distance between them is defined to be

$$d_L(A, B) = \inf\left\{ \left|\operatorname{dil}(f)\right| + \left|\operatorname{dil}\left(f^{-1}\right)\right| \right\},$$

where the infimum is over all homeomorphisms between A and B. If A and B are not homeomorphic, we define $d_L(A, B) = \infty$.

Note that for compact metric spaces A, B, $d_L(A, B) = 0$ if and only if A and B are isometric. It is not hard, then, to verify that d_L is a bona-fide metric on metric spaces (with possibly infinite value). The following proposition, whose proof can be found in [32], shows one connection between Lipschitz and Gromov–Hausdorff convergence.

PROPOSITION 74. *Suppose $\{Y_i\}$ is a sequence of compact metric spaces having uniformly bounded diameters. If Y is a metric space such that for every ε-net $\mathcal{N}$ in Y there exist ε_i-nets $\mathcal{N}_i$ in Y_i such that $\{\mathcal{N}_i\}$ converges to $\mathcal{N}$ in the Lipschitz metric and $\varepsilon_i \to \varepsilon$, then Y_i converges to Y in the Gromov–Hausdorff metric.*

The converse to the above proposition, that Gromov–Hausdorff convergence of Y_i to Y implies that nets in Y can be approximated by nets in Y_i in the Lipschitz metric, is also essentially true (cf. [32]).

We need to be able to study convergence of locally compact inner metric spaces, which need not be compact.

DEFINITION 75. A sequence (Y_i, p_i) of pairs, where Y_i is a locally compact inner metric space and $p_i \in Y_i$, is said to be (pointed) Gromov–Hausdorff convergent to a pair (Y, p) if for every $r > 0$, $\overline{B}(p_i, r)$ is Gromov–Hausdorff convergent to $\overline{B}(p, r)$.

Clearly pointed Gromov–Hausdorff convergence preserves the same metric properties as the nonpointed version, and the limit space Y is a locally compact inner metric space.

Finally, we give Gromov's Precompactness Criterion [32]:

THEOREM 76. *Let $\mathcal{X}$ be a collection of compact metric spaces. Then $\mathcal{X}$ is precompact (totally bounded) in the Gromov–Hausdorff metric if for every $\varepsilon > 0$ there exists an $N(\varepsilon)$ such that every $X \in \mathcal{X}$ has an ε-net with at most $N(\varepsilon)$ elements.*

4.3. *Products*

Given metric spaces (X_1, d_1), (X_2, d_2), we define the *product metric* on $X_1 \times X_2$ to be the metric $d((x_1, x_2), (y_1, y_2)) = \sqrt{d_1(x_1, y_1)^2 + d_2(x_2, y_2)^2}$. It is easy to see that (x_1, x_2) and (z_1, z_2) have a midpoint (y_1, y_2) if and only if y_1 is a midpoint of x_1 and z_1, and y_2 is a midpoint of x_2 and z_2. By adding ε's to the argument and applying Proposition 7 we easily obtain that the product of inner metric spaces is an inner metric space. Note that, if M_1 and M_2 are Riemannian manifolds and $M_1 \times M_2$ is given the product Riemannian metric, then the corresponding Riemannian distance in $M_1 \times M_2$ is the same as the metric product (in the above sense) of the Riemannian distances in M_1 and M_2. Furthermore, it is well-known in Riemannian geometry that $S_{k_1} \times S_{k_2}$ has sectional curvature $\geqslant \min\{0, k_1, k_2\}$. Therefore, if X_1 has curvature $\geqslant k_1$ and X_2 has curvature $\geqslant k_2$, then every quadruple in $X_1 \times X_2$ can be embedded in $S_{k_1} \times S_{k_2}$, and hence in some S_m with $m \geqslant \min\{0, k_1, k_2\}$. We have proved:

PROPOSITION 77. *If X_1, X_2 are spaces of curvature $\geqslant k_1, k_2$, respectively, then $X_1 \times X_2$ is a space of curvature $\geqslant \min\{0, k_1, k_2\}$.*

We next consider the space of directions at a point $(p, q) \in X_1 \times X_2$. It follows from our previous discussion of midpoints that a curve $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ is minimal if and only if $\gamma_1(t)$ and $\gamma_2(t)$ are minimal. More precisely, γ is a unit geodesic if and only if γ_i is unit for $i = 1, 2$, and $\gamma(t) = (\gamma_1(k_1 t), \gamma_2(k_2 t))$, where $k_1^2 + k_2^2 = 1$ and γ_1, γ_2 are unit; we write $\gamma = k_1\gamma_1 + k_2\gamma_2$. In other words, $\Sigma_{(p,q)}$ is the set of all linear combinations of elements of Σ_p and Σ_q whose coefficients are square summable to 1. By definition, the set $\Sigma_{(p,q)}$ is the join $\Sigma_p * \Sigma_q$. If $\gamma = k_1\gamma_1 + k_2\gamma_2 \in \Sigma_{(p,q)}$ and $\beta = m_1\beta_1 + m_2\beta_2 \in \Sigma_{(p,q)}$ then $\cos\alpha(\gamma, \beta) = k_1 m_1 \cos\alpha(\gamma_1, \beta_1) + k_2 m_2 \cos\alpha(\gamma_2, \beta_2)$. (This is the same formula one gets for the angle between two vectors in Euclidean space, written as a linear sum). From this formulation it is easy to see that if Σ_p and Σ_q are inner metric spaces, then so is $\Sigma_{(p,q)}$. We summarize the properties of $\Sigma_{(p,q)}$ in the following proposition:

PROPOSITION 78. *Let X_1, X_2 be spaces of curvature bounded below. For any $p \in X_1$, $q \in X_2$, $\Sigma_{(p,q)}$ consists of orthogonal, isometric copies of Σ_p and Σ_q, together with a unique segment of length $\pi/2$ joining each element of Σ_p to each element of Σ_q.*

Proposition 77 can be easily extended to finite products. Now suppose (X_i, d_i), $i = 1, 2, \ldots$, are metric spaces whose diameters are square summable. Then we can define the product metric on $X_1 \times X_2 \times \cdots$ by

$$d\big((x_1, \ldots), (y_1, \ldots)\big) = \sqrt{\sum_{i=1}^{\infty} d_i(x_i, y_i)^2}.$$

It is easy to see that $X = X_1 \times X_2 \times \cdots$ is the Gromov–Hausdorff limit of the finite products $X^k = X_1 \times \cdots \times X_k$, since each X^k isometrically embeds as a slice in X^{k+1}. Combining this with our earlier results in this section we obtain the following:

PROPOSITION 79. *Let (X_i, d_i) be spaces of curvature $\geqslant k$ whose diameters are square summable. Then $X_1 \times X_2 \times \cdots$ with the product metric is an inner metric space of curvature $\geqslant \min\{k, 0\}$.*

4.4. *Submetries, quotients*

DEFINITION 80. Let W, Y be metric spaces. A mapping $f : W \to Y$ is called a *weak submetry* (respectively *submetry*) if for each metric ball $B(x, r)$ (respectively closed metric ball $\overline{B}(x, r)$) in W, $f(B(x, r)) = B(f(x), r)$ (respectively $f(\overline{B}(x, r)) = \overline{B}(x, f(r))$.

The above definition is due to Berestovskii [11], and generalizes the notion of Riemannian submersion. Weak submetries are characterized by the property that for every $\varepsilon > 0$ and $y_1, y_2 \in Y$, there exist $x_i \in f^{-1}(y_i)$ such that $|d(x_1, x_2) - d(y_1, y_2)| < \varepsilon$. For strong submetries we can take $\varepsilon = 0$. The following is now not difficult to prove (cf. [16]):

PROPOSITION 81. *If $f : X \to Y$ is a weak submetry and X is an inner metric space of curvature $\geqslant k$, then Y is an inner space of curvature $\geqslant k$.*

Let G be a group of isometries of X. The quotient space X/G is the set of all equivalence classes $[x]$ of $x \in X$, were $x \equiv y$ if and only if there exists a $g \in G$ such that $g(x) = y$. The *orbit* of a point $x \in X$ is the set $G(x) = \{y\colon\ y \equiv x\} = \{g(x)\colon\ g \in G\}$. If each orbit $G(x)$ is closed, then for any $x, y \in X$, $d(x, G(y)) > 0$ if and only if $x \notin G(y)$. For any $z \in G(x)$, there is an isometry $g \in G$ such that $g(x) = z$. If $w \in G(y)$ and $d(x, w)$ is close to realizing $d(x, G(y))$, then $g(w) \in G(y)$ and $d(z, g(w)) = d(x, w)$. It follows that $d(x, G(y)) = d(z, G(y)) = d_H(G(x), G(y))$.

DEFINITION 82. Let G be a group of isometries of X such that the orbits of G are closed. We define the *quotient metric* on X/G by $d([x], [y]) = d_H(G(x), G(y))$.

Using the previous discussion it is not hard to verify that the above definition gives a *bona fide* metric. We also quickly obtain the following results:

PROPOSITION 83. *If G is a group of isometries of X such that the orbits of G are closed, then the natural quotient map $q : X \to X/G$ is a weak submetry when X/G has the quotient metric.*

COROLLARY 84. *Let G be a group of isometries of X such that the orbits of G are closed. Suppose that X/G is given the quotient metric.*

(1) *If X is an inner metric space then X/G is an inner metric space.*

(2) *If X has curvature $\geqslant k$ then X/G has curvature $\geqslant k$.*

REMARK 85. The orbits of G acting on X are closed if and only if the topological space X/G is Hausdorff.

4.5. *Cones, joins and simplicial complexes*

DEFINITION 86. Let X have diameter $\leqslant \pi$. The Euclidean cone on X is the quotient space $cX = X \times [0, \infty)/\equiv$, where $(x, t) \equiv (y, s)$ if and only if $s = t = 0$, with the following metric:

$$d\big((x, s), (y, t)\big) = \sqrt{s^2 + t^2 - 2st \cos d(x, y)}.$$

In the above definition we abuse notation slightly, to simplify the formulation. Another way to compute $d((x, s), (y, t))$ is to take two lines in the plane with angle $d(x, y)$ at 0, and measure length s to a point x_s on one line, and t to a point y_t on the other line. Then $d((x, s), (y, t)) = d(x_s, y_t)$. It follows from elementary trigonometry that if X is an inner metric space then cX is an inner metric space. For example, if x and y have a midpoint z in X, then (x, s) and (y, t), for $s, t > 0$ and $d(x, y) < \pi$, have a midpoint (z, r), where r is the distance from 0 to the intersection of the segment from x_s to y_t with a bisector of the angle between the lines. Note that if $k \geqslant 1$ then cS_k^n is a Euclidean cone with angle the diameter of S_k^n. We claim that cS_k^n has curvature $\geqslant 0$. To see this, let p denote the apex of the cone, and suppose, for simplicity, that $n = 1$. By slicing along a radial line from p and examining the behavior of straight lines in the resulting flat triangle, we see that no shortest curve between points away from p passes through p. Now consider a hinge $(\gamma_{ab}, \gamma_{ac})$ in cS_k^n. We can assume that $b \neq c$ (for then A2 is trivial). By continuity, we can assume that $c \neq p$. If we slice along a radial line from p through c, we see that $(\gamma_{ab}, \gamma_{ac})$ corresponds to a hinge $(\Gamma_{AB}, \Gamma_{AC})$ in the flat triangle, having equal side length and the same angle. However, since we are making identifications, $d(b, c) \leqslant d(B, C)$, A2 is satisfied.

Now a quadruple (a, b, c, d) in a X can be isometrically mapped onto a quadruple (a', b', c', d') in S_m for $m \geqslant 1$ if and only if the quadruple $((a, r), (b, s), (c, t), (d, u))$ in cX is isometric to the quadruple $((a', r), (b', s), (c', t), (d', u))$ in cS_m. Since cS_m is nonnegatively curved, we immediately obtain:

PROPOSITION 87. *X is an inner metric space of curvature $\geqslant 1$ if and only if cX is an inner metric space of curvature $\geqslant 0$.*

The following is an immediate consequence of the various definitions:

PROPOSITION 88. *If X is an inner metric space of curvature $\geqslant 1$ and p is the apex of cX, then $S_p = \Sigma_p$ is isometric to X.*

One problem with the Euclidean cone is that the subspace $(X, 1)$ is not generally a metrically embedded subspace. For this and other reasons it is useful to consider the *spherical cone* $c_1 X$, which is analogous to the Euclidean cone, except that it uses the cosine law for S_1.

DEFINITION 89. Let X have diameter $\leqslant \pi$. The spherical cone on X is the quotient space $c_1 X = X \times [0, \pi/2]/\equiv$, where $(x, t) \equiv (y, s)$ if and only if $s = t = 0$, with the following metric:

$$d\big((x, s), (y, t)\big) = \cos^{-1}\big(\cos t \cos s - \sin t \sin s \cos d(x, y)\big).$$

Note that the spherical cone on S_1^n is a hemisphere in S_1^{n+1}. For $k > 1$, we can use an argument analogous to the one above for the Euclidean cone, to see that $c_1 S_k^n$ has curvature $\geqslant 1$. For simplicity, we take $n = 1$. First note that if p is the "apex" of $c_1 S_k^1$, then at any point away from p, there is a neighborhood that is isometric to a neighborhood of a hemisphere H in S_1^2. In fact, we can take the neighborhood to be a "slice" $S(x, x')$ between radial geodesics from p to antipodal points x, x' on S_k^1. Thus the complement of $\{p\}$ has a lower curvature bound, so geodesics do not bifurcate. Now suppose that γ is a minimal curve starting at a point $a = (x, t) \neq p$. If γ is not a radial geodesic toward p, it cannot intersect p as it moves across a slice $S(x, x')$ (because this slice is isometric to a slice in H). Either γ runs into the boundary $S_k^1 \times \{\pi/2\}$ and stops, or it runs into the far side of the slice, to a point (x', s). Since there is another geodesic "going around the back" from (x, t) to (x', s), γ cannot continue beyond (x', s) as a minimal curve. In other words, any two points distinct from p are joined by a minimal curve not passing through p. To complete the proof as in the proof for the Euclidean cone, note that when we make a radial cut in $c_1 S_k^n$ we obtain a space isometric to H with a section between two radial geodesics removed, and apply the same argument. As above we obtain the following results:

PROPOSITION 90. *X is an inner metric space of curvature $\geqslant 1$ if and only if $c_1 X$ is an inner metric space of curvature $\geqslant 1$. Furthermore, X is metrically embedded in $c_1 X$ as the slice $X \times \{\pi/2\}$.*

PROPOSITION 91. *If X is an inner metric space of curvature $\geqslant 1$ and p is the apex of $c_1(X)$, then $S_p = \Sigma_p$ is isometric to X.*

If X has curvature $\geqslant 1$ then it is not hard to show that $c_1 X \cup_I c_1 X$, where $I : X \times \{\pi/2\} \to X \times \{\pi/2\}$ is the identity, is a space of curvature $\geqslant 1$, called the *spherical suspension* $\Sigma_1 X$. We now generalize this notion to that of the metric join:

DEFINITION 92. Let X, Y be inner metric spaces of curvature $\geqslant 1$. Let p be the apex of cX and q be the apex of cY. Then the *metric join* $X * Y$ of X and Y is defined to be the space of directions $S_{(p,q)}$ of $cX \times cY$.

The properties of the metric join can be seen in Proposition 78. We only emphasize that $X * Y$ is again a space of curvature $\geqslant 1$. If we take $Y = S^0$, then it is not hard to see that $X * Y$ is isometric to $\Sigma_1 X$. In addition, $S_1^n * S_1^m$ is isometric to S_1^{n+m+1}.

DEFINITION 93. Let X be an inner metric space of curvature $\geqslant 1$. Then the *q-fold spherical suspension* $\Sigma_1^q X$ is defined to be $S_1^{q-1} * X$, which is isometric to q iterated spherical suspensions $\Sigma_1 \Sigma_1 \cdots \Sigma_1 X$.

Clearly $\Sigma_1^q X$ is again a space of curvature $\geqslant 1$. There are related cone constructions (cf. [19, Section 4.3]). The *elliptic cone* produces a space of curvature $\geqslant -1$ from one of curvature $\geqslant 1$; the *parabolic cone*, a space of curvature $\geqslant -1$ from one of curvature $\geqslant 0$; and the *hyperbolic cone*, a space of curvature $\geqslant -1$ from one of curvature $\geqslant -1$.

A *simplex* in S_k^n is a convex subset of S_k^n homeomorphic to a standard topological simplex of dimension n. For example, a simplex in S_1 is a triangle (with its interior) whose sides are minimal curves, which is entirely contained in a closed hemisphere. Let C be a simplicial complex formed by gluing together finitely many simplices from a fixed space S_k^n, so that each codimension-1 face meets two simplices and the sum of the dihedral angles at any codimension-2 face is $\leqslant 2\pi$. Then C has curvature $\geqslant k$. Note that we have only to verify the curvature bound locally. First, C is locally a cone (of one of the above types) on another finite simplicial complex whose simplices are in S_1^{n-1}. The result can now be proved by induction on dimension.

5. Examples

EXAMPLE 94. Let M be a Riemannian manifold. Recall that this means that M is a smooth manifold, together with a smooth assignment of an inner product to the tangent space at each point. We measure the lengths of piecewise differentiable curves by integrating the lengths of its velocity vectors. The distance between two points is defined to be the infimum of the lengths of curves joining the points. It can be shown that this distance is inner. In fact, shortest curves are always geodesics in the sense of differential geometry: curves whose acceleration (in terms of the Riemannian connection) is 0. Equipped with this metric, a Riemannian manifold M is a space of curvature $\geqslant k$, if k is a lower bound for the sectional curvature of M. This follows from the Rauch Comparison Theorem (cf. [23]). Conversely, if M has curvature $\geqslant k$ in the sense of Definition 22, then M has sectional curvature $\geqslant k$. The space of directions S_p is always the unit sphere in the tangent space; M is metrically complete if and only if M is geodesically complete.

EXAMPLE 95. Let $\mathcal{R}(n, k, D)$ denote the space of all n-dimensional Riemannian manifolds having sectional curvature $\geqslant k$ and diameter $\leqslant D$. Then the closure $\overline{\mathcal{R}}(n, k, D)$ in the Gromov–Hausdorff metric consists of inner metric spaces having curvature $\geqslant k$ and

Hausdorff dimension $\leqslant n$. We remark that collapse is certainly possible: The flat torus $T^2 = S_1^1 \times S_k^1$ is Riemannian, with sectional curvature 0. By letting $k \to \infty$ we obtain a sequence of Riemannian manifolds with constant curvature 0 and bounded diameter converging in the Gromov–Hausdorff metric to S_1^1. For a discussion of collapse and convergence of Riemannian manifolds, see, e.g., [24]. If one smoothly rounds the apex of a Euclidean cone with angle $< \pi$, one can approximate the cone by Riemannian manifolds with nonnegative curvature. This shows that geometric singularities can arise in Gromov–Hausdorff convergence of Riemannian manifolds. It is also possible to construct a sequence of positively curved topological 2-spheres converging in the Gromov–Hausdorff metric to a line segment; thus topological singularities can also arise. Convergence of Riemannian manifolds will be discussed in greater detail later.

EXAMPLE 96. Let M be a closed Riemannian manifold of sectional curvature $\geqslant k$. and G be a compact group of isometries of M. We have already seen (Propositions 83 and 81) that M/G has curvature $\geqslant k$.

EXAMPLE 97. Every convex surface (see the introduction) is a space of curvature $\geqslant 0$. A particularly "bad" convex surface can be constructed in the following way (cf. [5]): Begin with the boundary of a simplex Σ_0 in $\mathbb{R}^3$. Take a barycentric subdivision of each face of Σ_0, and "push" the barycenter of each face outwards by a small amount, to obtain a new surface Σ_1. Subdivide Σ_1, and again push out the barycenters slightly – always doing so by an amount small enough that the surface remains convex, and no faces are at angle π with one another. Continuing this procedure, one obtains a space $\mathcal{B}$ such that the space of directions at a dense set of points has diameter strictly less than π – that is, there is a dense set of "singular" points. Note that this surface (and indeed all convex surfaces) can be approximated by polyhedra [2].

EXAMPLE 98. The Alexandrov's *double disk* Q is obtained by gluing two Euclidean disks along their boundaries. D can be obtained as a limit of convex surfaces, and so has curvature $\geqslant 0$. In fact, Q is usually considered as a kind of degenerate convex surface. The area of Q is maximal among all convex surfaces having the same diameter. It is a long-standing conjecture of Alexandrov that any convex surface of maximal area (relative to its diameter) must be isometric to Q. Note that Q is geodesically complete. In fact, geodesics are straight lines in the interior of either disk, then "flip" back and forth across the "rim" (identified boundaries) of Q. The rim itself is not a geodesic; it is shorter to follow one of the straight lines on either side, to join any pair of points on the rim. Q can also be approximated by Riemannian manifolds of positive curvature, spheres that are gradually flattened; the sectional curvature tends to $+\infty$ along the equator that becomes the rim in the limit.

EXAMPLE 99. Spaces of curvature bounded below that are definitely not manifolds or manifolds with boundary can be obtained by cones and suspensions, or quotients. For example, real projective space RP^n, being a covered by S^n, admits a Riemannian metric of constant sectional curvature 1. For $n > 1$, $\Sigma_1 RP^n$ has curvature $\geqslant 1$, but it is well known that this suspension is not a topological manifold.

EXAMPLE 100. Consider the Poincaré homology 3-sphere X, which, being covered by S^3, also admits a Riemannian metric of constant sectional curvature 1. Edwards' Double Suspension Theorem asserts that $\Sigma_1^2 X$ must be a manifold, and hence a sphere. By Propositions 91, 90, $\Sigma_1^2 X$ has curvature $\geqslant 1$, but at the two cone points the space of directions is isometric to $\Sigma_1^2 X$, which is definitely not a manifold, let alone a sphere. This example makes the important point that even when X is a topological manifold, the space of directions need not be a topological sphere.

EXAMPLE 101. Let $\{S_i\}$ be a sequence of circles of having square summable diameters. Then the metric product $T^\infty = S_1 \times S_2 \times \cdots$ is a compact abelian group with an inner metric of curvature $\geqslant 0$. Clearly the metric is *invariant* in terms of the group structure; that is, left (and right, since T^∞ is abelian) translation is an isometry. Note that T^∞ is infinite dimensional, but not an infinite dimensional manifold. T^∞ is a Gromov–Hausdorff limit of finite dimensional flat tori. The fundamental group of T^∞ is $Z \times Z \times \cdots$; in particular it has an infinite number of generators. This shows that dimension restrictions are required to bound the topology of a Gromov-Hausdorff sequence of spaces, even if the curvature is bounded above and below. Since any compact Lie group admits a bi-invariant metric of curvature $\geqslant 0$, this same construction can be carried out with any sequence of compact Lie groups. The resulting product group can have extremely complicated local topology. More generally, every locally compact, metrizable, arcwise connected topological group admits a compatible left-invariant metric of curvature $\geqslant k$ for some k [77,16]. These kinds of examples will be discussed further in Section 13.

6. Nonnegatively curved spaces

DEFINITION 102. A *ray* starting at $a \in X$ is an isometry $\rho : [0, \infty) \to X$ such that $\rho(0) = a$. A *line* (centered at a) in X is an isometry $\lambda : (-\infty, \infty) \to X$ such that $\lambda(0) = a$.

The condition that ρ or λ be an isometry is equivalent to ρ or λ being unit parameterized and minimizing when restricted to any finite subinterval. The next theorem is known as the Toponogov Splitting Theorem. It was proved by Toponogov for Riemannian manifolds, and for locally compact inner metric spaces of curvature $\geqslant 0$ in [39].

THEOREM 103. *Let X be a locally compact inner metric space of curvature $\geqslant 0$. If X contains a line λ then X splits as a metric direct product $X = \lambda \times Y$, where Y is a metrically embedded inner metric space of curvature $\geqslant 0$.*

To prove the theorem, let λ be a line centered at a in X. For any point $b \in X$ we construct a unique "line parallel to λ" through b, λ_b, by taking a sequence of minimal curves γ_i^+ from b to $\lambda(i)$, $i \in 1, 2, \ldots,$ which converges uniformly on every finite interval to a ray ρ^+ starting at b. Likewise we can construct a ray ρ^- as a limit of minimal curves γ_i^- from b to $-i$. We claim that ρ^+ and ρ^- fit together to form a line. In fact, consider the points $p := \rho^+(t)$ and $q := \rho^-(t)$. These are the limits of the points $\gamma_i^+(t)$ and $\gamma_i^-(t)$, respectively. Now basic Euclidean geometry shows that $\alpha_0(\gamma_i^+(t), \gamma_i^-(t)) \to \pi$. It follows from A0 that $d(p,q) = \lim d(\gamma_i^+(t), \gamma_i^-(t)) = 2t$. Uniqueness of λ_b follows from the fact that, if we were to choose another subsequence ω_j^+, we again obtain that, as i and j become large, $\alpha_0(\omega_j^+(t), \gamma_i^-(t)) \to \pi$, so $\alpha(\omega_j^+, \gamma_i^-) \to \pi$ and $\alpha(\omega_j^+, \gamma_i^+) \to 0$. We could apply the same argument to another subsequence ω_j^- to get uniqueness of the line.

Let Y denote the set of all $y \in X$ such that y is a closest point to a on the unique line λ_y through y parallel to a. Suppose that $z \in X$ is arbitrary, and let y be a point on λ_z closest to a. We claim that

$$d(a,z)^2 = d(a,y)^2 + d(y,z)^2. \tag{6}$$

If γ is minimal from y to a, $\alpha(\gamma, \lambda_z) = \pi/2$ by Lemma 64, and $d(a,z)^2 \leqslant d(a,y)^2 + d(y,z)^2$ by A2. To get the opposite inequality, suppose, for definiteness, that $z = \rho^+(t)$, where we are using the notation of the previous paragraph. Then $d(y,a) = \lim d(\gamma_i^+(t), a))$, and $d(\gamma_i^+(t), a))$ is, by A0, greater than or equal to the corresponding distance on the Euclidean triangle representing the triple of points $(\lambda(t); a, q)$. Since $\alpha_0(\lambda(t); a, q)) \to \pi/2$, the opposite inequality is obtained.

Formula (6) has two basic consequences:

(1) On each line parallel to λ, there is a unique point closest to a. Henceforth we will only denote a line parallel to λ by λ_b, where b is the point on λ_b closest to a.

(2) If λ_b is the line parallel to λ through b, then λ is the unique line parallel to λ_b through a, and a is the closest point to b on λ.

It now follows that formula (6) is valid if a is replaced by any point in Y. It remains only to show that Y is metrically embedded (Definition 66), for then it is immediate that Y is nonnegatively curved. Suppose that $a, y \in Y$ and let x be a midpoint between a and y. Then x lies on some λ_c, with $c \in Y$. If $c \neq x$ then $d(c,a) \leqslant d(x,a)$ and $d(c,y) \leqslant d(x,a)$, which violates the fact that x was a midpoint.

7. Analytical tools

In this section we present a few more technical tools needed to understand the structure of spaces of curvature bounded below. Throughout this section, X will denote a complete inner metric space of curvature $\geqslant k$, and $p \in X$.

7.1. *The tangent cone*

We now discuss the tangent cone, exponential map, and cut radius map.

DEFINITION 104. The *tangent cone* T_p at p is defined to be $c\Sigma_p$. The *exponential map* $\exp_p$ is defined to be the mapping from a subset of T_p into X defined by $\exp_p(t\gamma) = \gamma(t)$, where γ is a geodesic direction, and t is small enough that the definition makes sense.

REMARK 105. We will often refer to elements of T_p as "vectors" and adopt some vector notation. For $u \in T_p$, u is the equivalence class $[r, \gamma]$, where $\gamma \in S_p$ and $r \geqslant 0$, and this representation is unique unless $r = 0$. We write $u = r\gamma$, and define "scalar multiplication" by $cu = [cr, \gamma]$. We denote $[0, \gamma]$ simply by 0. We let $|u| := r = d(u, 0)$. For $u = r\gamma$, $v = c\beta \in T_p$ we let $\alpha(u, v) := \alpha(\gamma, \beta)$ and $|u - v|^2 = d(u, v)^2 = |u|^2 + |v|^2 - 2|u||v|\cos\alpha$. Finally, we let

$$\langle u, v\rangle = \frac{|u|^2 + |v|^2 - |u - v|^2}{2} = |u||v|\cos\alpha.$$

Certainly $\langle ru, v\rangle = \langle u, rv\rangle = r\langle u, v\rangle$ for any $r \geqslant 0$ and $|u|^2 = \langle u, u\rangle$. Note also that $\alpha(u, v)$ is the angle between the minimal curves tu and tv from 0 to u and v.

DEFINITION 106. Two vectors $u, v \in T_p$ are called *polar* (respectively *opposite*) if for all $w \in T_p$, $\langle u, w\rangle + \langle w, v\rangle \geqslant 0$ (respectively $\langle u, w\rangle + \langle w, v\rangle = 0$).

To see the geometric meaning of these terms, note that u and v are polar if and only if their directions are polar, so we can assume u and v are unit. Then if w is also unit, $\langle u, w\rangle + \langle w, v\rangle \geqslant 0$ is equivalent to $\alpha(u, w) + \alpha(w, v) \leqslant \pi$. In other words, every direction is within angle $\pi/2$ from either u or v. Two (possibly not unit) vectors u and v are opposite if and only if $|u - v| = 2|u| = 2|v|$, which is equivalent to $\alpha(u, v) = \pi$ (i.e., u and v are complementary in Σ_p) and u and v having the same length.

Note that by Proposition 87, if Σ_p is an inner metric space, then T_p is an inner metric space of curvature $\geqslant 0$. If $\Sigma_p = S_1^{n-1}$, then $T_p = \mathbb{R}^n$. In a Riemannian manifold, the exponential map is defined on all of $T_p = \mathbb{R}^n$, and is a diffeomorphism on some open set near the origin. In a space of curvature bounded below, the situation generally is not nearly so nice. To make this precise we need a few geometric notions.

DEFINITION 107. The *cut radius map* is the mapping $C_p : \Sigma_p \to \mathbb{R}^+ \cup \infty$ given by $C_p(\gamma) = \sup\{t\colon \gamma|_{[0,t]}$ is defined and minimal$\}$ if $\gamma \in S_p$ and $C_p(\gamma) = 0$ otherwise.

The cut radius map measures how far out a geodesic extends as a minimizing curve. There are two possibilities: The geodesic continues past the cut radius as a geodesic (which always happens in the geodesically complete case, or it stops, as does a geodesic that hits the apex of a Euclidean cone (with angle $< \pi$).

DEFINITION 108. A point $p \in X$ is called a *geodesic terminal* if there exists a geodesic γ_{qp} which cannot be extended past p.

Note that X is geodesically complete if and only if it has no geodesic terminals. It is instructive to consider the simple case of Q (Example 98). If p is a point on the rim, then T_p is isomorphic to the Euclidean plane. Then $\exp_p$ is defined on the plane with a line (say, the x-axis) removed, except for the origin. The removed half-lines point in directions "tangent" to the rim. A geodesic starting at p is minimizing until it strikes the rim again; that is where the cut radius occurs, since another shortest path can be found by travelling along the other side of Q. In this case the cut radius is continuous.

It is not hard to see that, in general, the cut radius map is semicontinuous. For a Riemannian manifold it is continuous, but in general it is not. For example, the product space $S^1 \times Q$ is geodesically complete, but C_p is not continuous in the directions tangent to an S^1 slice. In fact, it is possible for C_p to have no points of continuity on S_p (it is always continuous at nongeodesic directions):

EXAMPLE 109. Consider the "bad sphere" $\mathcal{B}$ of Example 97. This convex surface has a dense set of singular points, each of which is a geodesic terminal; more strongly, no geodesic passes through such a singular point. We claim that there is no function $f : \Sigma_p \to \mathbb{R}^+$ such that $f \leqslant C_p$, $f > 0$ on S_p and f is continuous at some $\gamma \in S_p$. If there were such a function, with $f(\gamma) = \varepsilon > 0$, then any geodesic α near γ (in angle) would satisfy $C_p(\alpha) \geqslant f(\alpha) > \varepsilon/2 > 0$. But then there would be no singular points near $\gamma(\varepsilon/2)$, since all points near $\gamma(\varepsilon/2)$ would have at least one geodesic passing through them. This contradicts the fact that the singular points are dense. It follows that, in this case, the domain of $\exp_p$ does not contain an open set. The fact that the domain of $\exp_p$ may not contain an open set makes it much more difficult to understand the local structure of spaces of curvature $\geqslant k$, because $\exp_p$ is not a local homeomorphism as it is for Riemannian manifolds.

7.2. *Concave functions*

We now recall a few elementary facts from analysis. A function $f : Y \to Z$ between metric spaces is said to be *λ-Lipschitz* if for every $x, y \in Y$, $d(f(x), f(y)) \leqslant \lambda d(x, y)$. The number λ is called the Lipschitz constant for f. When it is unimportant, we do not mention λ specifically. A Lipschitz function is obviously uniformly (and absolutely) continuous. If $Y = [a, b]$ and $Z = [c, d]$ then f, being absolutely continuous, is differentiable almost everywhere (a.e.), and $f(x) = f(a) + \int_a^x f'(t)\,dt$. A Lipschitz curve c in a metric space is obviously rectifiable, with $L(c) \leqslant \lambda(b - a)$. The next lemma is an immediate consequence of the triangle inequality.

LEMMA 110. *If $c : [a, b] \to Z$ is λ-Lipschitz then for any $p \in Z$, the function $f(t) = d(p, c(t))$ is λ-Lipschitz from $[a, b]$ into $\mathbb{R}^+$.*

Given any rectifiable curve $c : [a, b] \to Y$, the length $L(t) = L(c|_{[a,t]})$ is monotone increasing, and so differentiable a.e. Then $L(c) = \int_a^b L'(t)\,dt$. For any $t \in (a, b)$,

$$d\big(c(t), c(t + \delta)\big) \leqslant \big|L(t + \delta) - L(t)\big|$$

for any $\delta > 0$ near 0, and so the limit $\lim_{\delta \to 0} d(c(t), c(t + \delta))/|\delta|$ exists a.e.; we denote this limit by $|c'(t)|$. It follows from the definition of length that $L(c) = \int_a^b |c'(t)|\,dt$. Note that if c is unit, then $|c'(t)| = 1$ a.e. If f is Lipschitz and 1–1 and f^{-1} is also Lipschitz, we say that f is *bi-Lipschitz*; clearly f is a homeomorphism in this case.

Recall that a function $f : [a, b] \to \mathbb{R}$ is called *concave* if the segment between any pair of points on the graph of f lies below (not necessarily strictly) the graph. The following statements are equivalent:

(1) f is concave,
(2) f is continuous and for any x, y, $f(\frac{x+y}{2}) \geqslant \frac{f(x)+f(y)}{2}$, and
(3) f is the integral of a monotone decreasing function f'.

If f is concave then the secant slope $\frac{f(s)-f(t)}{s-t}$ is a decreasing function of both s and t $(s < t)$. It follows that the left and right derivatives

$$f^{-}(t) = \lim_{s \to t^{-}} \frac{f(s) - f(t)}{s - t} \quad \text{and} \quad f^{+}(t) = \lim_{s \to t^{+}} \frac{f(s) - f(t)}{s - t}$$

exist for all t, and are approached in a monotone way. (Since f is Lipschitz, $f^{-}(t) = f^{+}(t)$ a.e.) We now adopt the notation of [65].

DEFINITION 111. Let $\phi, f, F : [a, b] \to \mathbb{R}$ be continuous functions such that $F'' = f$. We write $\phi'' \leqslant f$ if $\phi - F$ is concave.

If ϕ is twice differentiable, then $\phi'' \leqslant f$ has two (equivalent) meanings. Note that a function ϕ is concave if and only if $\phi'' \leqslant 0$. Also, if $\phi'' \leqslant f$ then a.e. ϕ' exists and is equal to the sum of a monotone decreasing function $h(t)$, and the differentiable function F'. For the convenience of the reader we state a few basic results connected with Definition 111. These follow from the basic properties of concave functions that were discussed above, plus elementary facts like: concave functions are closed under certain algebraic operations and inf's.

LEMMA 112. *Let $\phi, \phi_i, f, F, g, G : [a, b] \to \mathbb{R}$ be continuous functions such that $F'' = f$ and $G'' = g$. Then*

(1) *If $\phi_1'' \leqslant f$ and $\phi_2'' \leqslant g$ then $(\phi_1 + \phi_2)'' \leqslant f + g$.*
(2) *If $\phi'' \leqslant f$ and $c > 0$ then $(c\phi)'' \leqslant cf$.*
(3) *If $\phi'' \leqslant f$ and $f \leqslant g$ then $\phi'' \leqslant g$.*
(4) *If $\phi_i \to \phi$ on $[a, b]$ and $\phi_i'' \leqslant f$ for all i, then $\phi'' \leqslant f$.*
(5) *If $\phi_i'' \leqslant f$ for all i, then $(\inf \phi_i)'' \leqslant f$.*
(6) *If $\phi'' \leqslant f$ and g is concave, then $(g \circ \phi)'' \leqslant \sup\{g'\} f$.*
(7) *If $\phi'' \leqslant f$ then $\phi^{-}(t)$ and $\phi^{+}(t)$ exist for all t and are equal a.e.*
(8) *If $\phi'' \leqslant f$ then $\phi^{+}(t) \geqslant \frac{\phi(s)-\phi(t)}{s-t} - \max_{\tau \in [t,s]} |f(\tau)|(s - t)$ whenever $s > t$.*
(9) *Finally, $\phi'' \leqslant f$ if and only if $(\phi + G)'' \leqslant f + g$.*

LEMMA 113. *If $\phi'' \leqslant f$ on $[a, c]$ and $[c, b]$ for some $c \in (a, b)$ then $\phi'' \leqslant f$ on $[a, b]$ if and only if $\phi^{-}(c) \geqslant \phi^{+}(c)$.*

The above lemma allows us to "glue" together concave functions.

If $\phi'' \leqslant f$, then for any fixed t we can write, for some constant K,

$$\phi(t + \tau) \leqslant \phi(t) + K\tau + \int_t^{t+\tau} \int_t^{s} h'(\sigma) + f(\sigma) \, d\sigma \, ds.$$

Here h is the decreasing function mentioned above Lemma 112; we immediately get

$$\phi(t+\tau) \leqslant \phi(t) + K\tau + \frac{f(t)}{2}\tau^2 + \frac{\tau^2}{2} \max_{|s-t|\leqslant\tau} \{|f(s) - f(t)|\},$$

where the last term is $o(\tau^2)$, since f is continuous. The converse statement is easy, and we have:

LEMMA 114. *For any $\phi, f : [a,b] \to \mathbb{R}$, $\phi'' \leqslant f$ if and only if for some constant K,*

$$\phi(t+\tau) \leqslant \phi(t) + K\tau + \frac{f(t)}{2}\tau^2 + o(\tau^2).$$

7.3. *Development of curves and quasigeodesics*

We have seen (and Example 109 emphasizes) that the usefulness of geodesics is more limited for general spaces of curvature $\geqslant k$ than it is for Riemannian manifolds. Quasigeodesics were introduced by Alexandrov [1] in order to overcome some of those limitations. We use here an "invariant" definition that comes from Proposition 1.7 in [65], and does not need a curvature bound.

DEFINITION 115. Let Y be a metric space and $\eta : [a,b] \to Y$ be a unit curve. Then η is called a *quasigeodesic* if for every $q \in Y$ there exists a function f such that $\lim_{x\to 0} \frac{f(x)}{x} = 0$, and $(d_q(t)^2 - t^2)'' \leqslant f(d_q(t))$ for all $t \in [a,b]$. We write $(d_q(t)^2 - t^2)'' \leqslant o(d_q(t))$ for short.

Our next task is to show that quasigeodesics, while generally not geodesics (as we will see later), have the same "comparison" properties. If p is a point in Euclidean space, γ is a line, and α is the angle between γ and the line from p to $\gamma(t)$ then letting $d_s := d(p, \gamma(s))$, we can write $d_{t+\tau}^2 = d_t^2 - (2d_t \cos\alpha)\tau + \tau^2$. In a space of curvature $\geqslant 0$, we know that the corresponding distance $\tilde{d}_{t+\tau}$ is smaller than $d_{t+\tau}$ (if $d_t = \tilde{d}_t$ and we use the same angle α), and we can write $\tilde{d}_{t+\tau}^2 \leqslant \tilde{d}_t^2 - (2\tilde{d}_t \cos\alpha)\tau + \tau^2$. Using the same arguments and the cosine laws for space forms, one can prove:

PROPOSITION 116. *Let $p \in X$ and γ be a geodesic in X, and let $d_t = d(p, \gamma(t))$. Let*

$$f(t) = \begin{cases} \frac{1-\cos(\sqrt{k}d_t)}{k} & \text{if } k > 0, \\ \frac{d_t^2}{2} & \text{if } k = 0, \\ \frac{1-\cosh(\sqrt{-k}d_t)}{k} & \text{if } k < 0. \end{cases}$$

Then $f'' \leqslant 1 - kf$.

Conversely, a locally compact inner metric space in which the conclusion of Proposition 116 always holds (for fixed k), for every geodesic, must have curvature $\geqslant k$. To see this in the case when $k = 0$, consider the expression $d_{t+\tau}^2 \leqslant d_t^2 + K\tau + \tau^2 + o(\tau^2)$. Argue that X must have some lower curvature bound k; this determines the value of $K = 2d_t \cos\alpha$, and the result follows. However, we will not need this result.

In S_k we fix a point p; the space of directions Σ_p is a circle of diameter π; we fix a monotone reparameterization $C(\theta)$ of Σ_p (i.e., $C(\theta) = (\cos(f(\theta)), \sin(f(\theta))$ for some monotone surjection $f:\mathbb{R} \to \mathbb{R}$). In other words, $C(\theta)$ moves, say, counterclockwise, without turning back. Then every point $q \in S_k$ has *C-polar coordinates* $(r, \theta)_C$, where $r = d(p, q)$, and θ is such that $C(\theta)$ is the direction of a minimal curve from p to q. If $0 < d(p, q) < \pi/\sqrt{k}$ then θ is uniquely defined in the interval $f^{-1}([0, 2\pi))$. Polar coordinates always uniquely determine a point, although they are not defined when $r > \pi/\sqrt{k}$. Note that a curve $c(t) = (r(t), \theta(t))$ is rectifiable if and only if each of $r(t)$ and $\theta(t)$ is rectifiable. When f is the identity map, we simply write (r, θ) for $(r, \theta)_C$ (these are the standard polar coordinates in Euclidean space).

DEFINITION 117. Let Y be a metric space and $c:[a, b] \to Y$ be a Lipschitz curve and $p \in Y$. The k-development of c at p is the curve $\tilde{c}:[a, b] \to S_k$ whose C-polar coordinates are $(d(p, c(t)), t)_C$, where the parameterization C is chosen so that $\tilde{c}$ is unit parameterized.

From Lemma 110 and the above comments, we see that the curve $(d(p, c(t)), t)$ is at least rectifiable; we can then choose a unique mapping f to make $(d(p, c(t)), t)_C$ unit parameterized (cf. [65]). That is, as long as $d(p, c(t)) < \pi/\sqrt{k}$, we have a unique (up to isometric motion of S_k) unit k-development of c at p; we implicitly assume $d(p, c(t)) < \pi/\sqrt{k}$ when discussing the k-development in the future. The k-development of a curve was discovered by Alexandrov [1]. The k-development $\tilde{c}$ of a curve c is called *convex* if the following holds for any t and sufficiently small $\varepsilon > 0$. Let $x = (r, \theta)_C$ lie on the minimal curve γ joining $\tilde{c}(t)$ and $\tilde{c}(t + \varepsilon)$, with $t < \theta < t + \varepsilon$. Then $r \leqslant d(p, c(\theta))$. In other words, the "section" determined by $\tilde{c}$ in $[t, t + \varepsilon]$ is convex in S_k. (The word "convex" is greatly over-used; in this case we say a subset S of an inner metric space is convex if every pair of points in S is joined by a minimal curve lying in S.) We now have yet another useful way to characterize curvature $\geqslant k$. We give only one direction here; if one assumes local compactness (or the existence of sufficiently many minimal curves), then one obtains the converse. The proof of this theorem is not hard. In fact, the convexity of a k-development is equivalent to geodesics satisfying the conditions of Proposition 116.

THEOREM 118. *For any $p \in X$ and geodesic $\gamma : [a, b] \to X$, the k-development of γ at p is convex.*

DEFINITION 119. If Y is a metric space, a curve $\eta:[a, b] \to Y$ is called *k-convex* if for every $p \in Y$, the k-development of η at p is convex.

Note that a curve c which is k-convex is k'-convex for any $k' \leqslant k$.

PROPOSITION 120. *A curve c in X is a quasigeodesic if and only if it is k-convex and unit parameterized.*

The "only if" direction follows from the formulas in Proposition 116, and Lemma 114. (Use the Maclaurin's series when $k \neq 0$.) The converse involves a little more analysis (cf. [65], Proposition 1.7).

7.4. *Differential and gradient*

Yet another kind of curve that will be of use to us is gradient curves. We give a slightly modified version of the definitions and results of [65]. Recall that the gradient ∇f of a smooth function f on a Riemannian manifold measures the direction and magnitude maximal increase of f. Such a direction exists due to the smoothness of f and the compactness of the space of directions at every point. We begin with the following:

DEFINITION 121. Let Y be a metric space and $f: U \subset Y \to \mathbb{R}$ be λ-Lipschitz, where U is open. The *absolute gradient* of f at $p \in U$ is defined to be

$$|\nabla f|(p) = \max\left\{0, \lim \sup_{p_i \to p} \frac{f(p_i) - f(p)}{d(p, p_i)}\right\} \leqslant \lambda.$$

A point p is called *critical* for f if $|\nabla f|(p) = 0$.

DEFINITION 122. For any $A \subset X$ and $p \in X$, we denote by γ_{pA} a minimal curve from p to a point in A such that $L(\gamma_{pA}) = d(p, A) := \min_{x \in A} d(p, x)$. Let A'_p denote the subset of Σ_p corresponding to curves γ_{pA}. When the reference to p is clear we will eliminate the subscript to simplify notation. When $A = \{q\}$, we will simply write q'.

Note that A' is always closed. It follows easily from A1 that, for the distance function d_A, the above notion of critical point is equivalent to the extremely useful one introduced in [44] for Riemannian manifolds (it can be stated in the current generality).

PROPOSITION 123. *Let A be a subset of X. A point q is a critical point of d_A if and only if A'_q forms a $\pi/2$-net in Σ_q.*

In other words, it is impossible to move from q in a direction that points away from A in the sense that its angle with some minimal curve from q to A is more than $\pi/2$. As may be expected the notion of absolute gradient is only useful with functions having fairly strong properties.

DEFINITION 124. Let Y be an inner metric space. A Lipschitz function $f: U \subset Y \to \mathbb{R}$ is said to be κ*-concave* if for every geodesic γ in U, $(f \circ \gamma)'' \leqslant \kappa$. The function f is called *semiconcave* if for every $p \in U$ there exists a $\kappa(p) > 0$ such that f is $\kappa(p)$ concave near p.

Note that our definition differs by a constant multiple from the definition of "λ-convexity" in [65]. It follows from Lemma 112 that a function is κ-concave if and only if there exists a bounded function g such that for every geodesic γ in U, $(f \circ \gamma)'' \leqslant g$.

Note that for $k=0$, the formula of Proposition 116 is $(d_t^2)'' \leqslant 2$. By using the Maclaurin's expansion of $\cos t$ and $\cosh t$ and Lemma 114 one can show that d_t^2 is semiconcave for $k \neq 0$ as well, where the local parameter κ depends continuously on the distance from p. Since $d_A = \inf_{q\in A}\{d_q\}$, we get from Lemma 112 the following proposition.

PROPOSITION 125. *For any closed $A \subset X$, d_A is semiconcave on $X\backslash A$, and d_A^2 is semiconcave on X.*

We will need a technical lemma, which follows from the definition of concavity (of $\frac{\kappa t^2}{2} - f$):

LEMMA 126. *Suppose f is a real function such that $f'' \leqslant \kappa$ for some constant κ. Then for any $0 < s < t$, $f(s) > f(0) + \frac{s}{t}(f(t) - f(0)) - |\kappa| s(t-s)$.*

DEFINITION 127. If Y is a metric space, $f: U \subset Y \to \mathbb{R}$ is semiconcave and γ is a unit geodesic starting at p, $p \in Y$, we denote $(f \circ \gamma)^+$ by $D_\gamma(f)$.

The number $D_\gamma(f)$ is the "directional derivative" of f in the direction of γ. The following proposition is immediate from Lemma 33:

PROPOSITION 128. *Let $f: X \to \mathbb{R}$ be semiconcave with Lipschitz constant λ. Then for any $\gamma_1, \gamma_2 \in S_p$, $p \in X$,*

$$\left|d_{\gamma_1}(f) - d_{\gamma_2}(f)\right| \leqslant 2\lambda \sin \frac{\alpha(\gamma_1, \gamma_2)}{2}.$$

In particular, the mapping $D_*(f): S_p \to \mathbb{R}$ is continuous and has a unique continuous extension to Σ_p and also to T_p.

DEFINITION 129. We denote the continuous extension of $D_*(f)$ to T_p by df_p, or simply df; it is called the *differential* of f at p.

Obviously, $|\nabla f|(p) \geqslant \sup\{D_*(f)\}$. From Lemma 112 we get the first part of the following proposition. The second part follows from the first part and Theorem 27.

PROPOSITION 130. *If Y is a metric space, $f: U \subset Y \to \mathbb{R}$ is κ-concave, and γ is a geodesic starting at $p \in Y$, then*

$$|\nabla f|(p) \geqslant D_\gamma(f) \geqslant \frac{f(\gamma(t)) - f(p)}{t} - |\kappa| t$$

whenever $t > 0$ is small enough that γ restricted to $[0, t]$ lies in U. Suppose, furthermore, that Y is a complete inner metric space of curvature $\geqslant k$ and U is a metric ball $B(p, r)$, $r > 0$. Then for all $q \in U$,

$$|\nabla f|(p) \geqslant \frac{f(q) - f(p)}{d(p, q)} - |\kappa| d(p, q).$$

COROLLARY 131. *If $f : U \subset X \to \mathbb{R}$ is semiconcave then $|\nabla f|$ is semicontinuous; that is,* $\liminf_{p_i \to p} |\nabla f|(p_i) \geqslant |\nabla f|(p)$.

COROLLARY 132. *If $f : U \subset X \to \mathbb{R}$ is semiconcave then the set of all noncritical points of f is open in U.*

COROLLARY 133. *If $f : U \subset X \to \mathbb{R}$ is semiconcave then the function d_p has no critical points in a neighborhood of $p \in U$.*

COROLLARY 134. *If $f : U \subset X \to \mathbb{R}$ is semiconcave then* $|\nabla f|(p) = \sup\{D_*(f)\}$.

To prove Corollaries 131 and 134, fix positive $\varepsilon < 1/2$ and choose p' such that

$$\frac{f(p') - f(p)}{d(p, p')} > |\nabla f|(p) - \varepsilon$$

and $d(p, p') < \varepsilon$. Let q be close enough to p that $f(p') - f(q) \geqslant (1 - \varepsilon)(f(p') - f(p))$ and $d(p', q) \leqslant (1 + \varepsilon)d(p', p) < 2\varepsilon$. Then by Proposition 130 (choosing κ by the semiconcavity),

$$\begin{aligned} |\nabla f|(q) &\geqslant \frac{f(p') - f(q)}{d(p', q)} - |\kappa| d(p', q) \geqslant \frac{1 - \varepsilon}{1 + \varepsilon}\left(\frac{f(p') - f(p)}{d(p, p')}\right) - 2|\kappa|\varepsilon \\ &> \frac{1 - \varepsilon}{1 + \varepsilon}|\nabla f|(p) - \left(2|\kappa| + \frac{1}{3}\right)\varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$ proves the first corollary. For the latter one, by moving p' slightly, we can assume there is a minimal curve γ from p to p'. Again apply Proposition 130 to get $D_\gamma(f) \geqslant |\nabla f|(p) - (1 + |\kappa|)\varepsilon$. Let $\varepsilon \to 0$.

DEFINITION 135. Suppose that $f : U \subset X \to \mathbb{R}$ is κ-concave. We make the following definition for any $p, q \in X$:

$$\Delta(p, q) := \frac{f(q) - f(p)}{d(p, q)} - |\kappa| d(p, q).$$

PROPOSITION 136. *Suppose that $f : U \subset X \to \mathbb{R}$ is κ-concave, and U is bounded. For every $\varepsilon > 0$ there exists a $\delta(k, \kappa)$ such that for any p, q_1, q_2, if*

$$\Delta(p, q_i) > (1 - \delta)|\nabla f|(p)$$

for $i = 1, 2$, then $\alpha_k(p; q_1, q_2) < \varepsilon$.

This proposition says essentially that points which come close to realizing the gradient must be nearly "lined up". We prove it as follows: First, by moving the points q_i slightly we can assume that there are unit minimal curves $\gamma := \gamma_{pq_1}$ and $\beta := \gamma_{pq_2}$, as well as unit minimal curves β_i from $x_i := \gamma(2^{-i})$ to q_2. Setting $y_i := \beta_i(2^{-i})$, we obtain from Lemma 126 that

$$f(y_i) - f(x_i) \geqslant \frac{d(x_i, y_i)}{d(x_i, q_2)}\big(f(q_2) - f(x_i)\big) - |\kappa| d(x_i, y_i) d(y_i, q_2).$$

If the formula in Proposition 136 holds for some δ, then for i large enough we get

$$\frac{(f(q_2) - f(x_i))}{d(x_i, q_2)} > (1 - 2\delta)|\nabla f|(p) + |\kappa| d(x_i, q_2)$$

and then

$$\begin{aligned} f(y_i) - f(x_i) &> d(x_i, y_i)(1 - 2\delta)|\nabla f|(p) + |\kappa| d(x_i, y_i)^2 \\ &> d(x_i, y_i)(1 - 2\delta)|\nabla f|(p). \end{aligned}$$

Likewise,

$$f(x_i) - f(p) > d(p, x_i)(1 - 2\delta)|\nabla f|(p).$$

Now from Proposition 130 we get

$$d(p, y_i) \geqslant \frac{1}{|\nabla f|(p)}\big(f(y_i) - f(p) - |\kappa| d(p, y_i)^2\big).$$

Adding and subtracting $f(x_i)$ to the right side of the last inequality, and combining it with the previous two inequalities gets us

$$d(p, y_i) > (1 - 2\delta)\big(d(p, x_i) + d(x_i, y_i)\big) - \frac{|\kappa|}{|\nabla f|(p)} d(p, y_i)^2.$$

From Lemma 16 it now follows that $\alpha_k(x_i; y_i, p)$ is close to π, so $\alpha_k(x_i; q_2, p)$ is close to π, and $\alpha_k(x_i; q_2, q_1)$ is close to 0. Letting $x_i \to p$ completes the proof. The next corollary was proved in the finite dimensional case in [65].

COROLLARY 137. *Suppose that $f : U \subset X \to \mathbb{R}$ is κ-concave and without critical points in U. Then for every $p \in U$ there exists a unique direction $\gamma \in \Sigma_p$ such that $|\nabla f|(p) = df(\gamma)$.*

PROOF. Suppose that $\gamma_1, \gamma_2 \in S_p$ satisfy $df(\gamma_i) > (1 - \delta/4)|\nabla f|(p)$ for some $\delta > 0$. Then for $t > 0$ sufficiently small,

$$\frac{f(\gamma_i(t)) - f(p)}{t} > \left(1 - \frac{\delta}{2}\right)|\nabla f|(p) > (1 - \delta)|\nabla f|(p) + |\kappa| t.$$

Given $\varepsilon > 0$, Proposition 136 now implies that for δ small enough, $\alpha_k(p; \gamma_1(t), \gamma_2(t)) < \varepsilon$ for all small t. In other words, $\alpha(\gamma_1, \gamma_2) < \varepsilon$, and the corollary follows. □

DEFINITION 138. We denote by $\nabla f(p) \in T_p$ the vector $|\nabla f|(p) \cdot \gamma$, where γ is the direction given by the above corollary, if $|\nabla f|(p) > 0$; otherwise we let $\nabla f(p) := 0$. The vector $\nabla f(p)$ is called the *gradient* of f at p.

If f has a local maximum at p then Definition 121 obviously implies that $\nabla f(p) = 0$. If $\nabla f(p) > 0$ then we can find a geodesic direction γ such that $df_p(\gamma) > 0$, and so f does not have a local minimum at p. We have proved:

PROPOSITION 139. *Suppose that $f : U \subset X \to \mathbb{R}$ is κ-concave. Then $p \in U$ is extremal (i.e., is either a local max or a local min) only if $\nabla f(p) = 0$.*

7.5. *Gradient curves*

The construction which follows is roughly from [65], which in turn is similar to that given in [79, Theorem 3].

DEFINITION 140. Let Y be a metric space and $f : U \subset Y \to \mathbb{R}$ be a Lipschitz function. A curve $c : (a, b) \to U$ is called an *f-gradient curve* if

$$\lim_{s \to t^+} \frac{f(c(s)) - f(c(t))}{d(c(s), c(t))} = |\nabla f|\big(c(t)\big).$$

That is, an f-gradient curve is one along which one can measure the gradient of f. Note that being an f-gradient curve is independent of (orientation preserving, monotone) reparameterization. If U has no critical points of f, then the function $h(t) = f(c(t))$ is strictly increasing. If we reparameterize, letting $b(t) := c(h^{-1}(t))$ then we get that $f(b(t)) = t$. A curve b satisfying $f(b(t)) = t$ will be called *f-parameterized*. Clearly the f-parameterization is unique. The next proposition, whose proof is now easy, reconciles our definition with the one in [65].

PROPOSITION 141. *Let Y be a metric space and $f : U \subset Y \to \mathbb{R}$ be a Lipschitz function without critical points. A curve $c : (a, b) \to U$ is an f-gradient curve if and only if c has an f-parameterization such that*

$$\lim_{s \to t^+} \frac{d(b(s), b(t))}{s - t} = \frac{1}{|\nabla f|(b(t))}.$$

Let us pause for a moment to see what these definitions mean in a Riemannian manifold, in the case when c and f are smooth and f is without critical points (i.e., ∇f never vanishes) in U. To compute $|\nabla f|(p)$ in the Riemannian sense we can choose a geodesic γ such that $\gamma'(0) = \frac{\nabla f(p)}{|\nabla f(p)|}$ and compute

$$\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(0) = \lim_{s \to t} \frac{f(\gamma(s)) - f(\gamma(t))}{s - t} = \lim_{s \to t} \frac{f(\gamma(s)) - f(\gamma(t))}{d(\gamma(s), \gamma(t))}.$$

On the other hand, for any $p_i \to p$, so that $\lim_{i \to \infty} \frac{f(p_i) - f(p)}{d(p, p_i)}$ exists and equals some number L, then by choosing a subsequence if necessary we can assume that the directions γ_{pp_i} converge to $\gamma \in S_p$. We see that for some $t_i \to 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(0) = \lim_{t_i \to 0} \frac{f(\gamma(t_i)) - f(\gamma(0))}{t_i} = L.$$

It follows that the two notions of $|\nabla f|$, and hence of critical point, coincide. Now let c be such that $c'(t) = \nabla f(c(t))$. Then for some real function g, $f(c(g(t))) = t$ if and only if $g'(t) = 1/|\nabla f|^2(c(g(t)))$, that is,

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} c\big(g(t)\big)\right| = \frac{1}{|\nabla f|(c(g(t)))}.$$

Now

$$\lim_{s \to t^+} \frac{s - t}{d(c(g(s)), c(g(t)))} = \lim_{s \to t^+} \frac{f(c(g(s))) - f(c(g(t)))}{d(c(g(s)), c(g(t)))} = |\nabla f|\big(c\big(g(t)\big)\big).$$

In other words, c is an f-gradient curve, and every f-gradient curve is simply a reparameterization of c.

PROPOSITION 142. *Let f be semiconcave and without critical points in $U \subset X$. Then for each $p \in U$ there exists a unique (up to reparameterization), complete gradient curve starting at p.*

A proof of the above proposition is sketched in the appendix of [65]. It is based on the following extension of Proposition 136.

PROPOSITION 143. *Let f be κ-convex and without critical points in U. Then there exists a constant C such that for any sufficiently small $\delta > 0$ the following holds: Let $p, a, b \in U$ such that $\Delta(p, a) \geqslant (1 - \delta^2)|\nabla f|(p)$. Then*

$$\cos \alpha_k(p; a, b) \geqslant \frac{\Delta(p, b)}{|\nabla f|(p)} + C\delta.$$

To construct a gradient curve, one first constructs a kind of approximating gradient sequence. (We roughly follow the sketch in [65], with a few more details in some spots.) Fix a small $\delta > 0$. Begin with p, choose a point q such that $d(p, q) < \delta$ and

$\Delta(p,q) \geqslant (1-\delta^2)|\nabla f|(p)$ for some small $\delta > 0$. If the construction leads to an accumulation point, we begin a new sequence. An application of Zorn's Lemma allows us to continue this process until we no longer have an accumulation point in U. At this point we have constructed a possibly infinite sequence indexed over a well-ordered subset of the rationals which itself has possibly countably many accumulation points. Suppose x_α is such an accumulation point. We choose a point x_β with $\beta < \alpha$ very close to x_α, and remove all points with indices between α and β. We call x_β and exceptional point. We carry this procedure out at (countably many) accumulation points so that the sum of the distances between exceptional points and their successors is finite and small. We re-index the sequence over the integers. Given two such sequences $\{x_i\}$ and $\{y_i\}$ we can estimate rate of change of the distance between two points x_i and y_j such that $0 \leqslant f(y_j) - f(x_i) < \delta^2$, and x_i and y_j are not exceptional. By the above proposition,

$$\begin{aligned}\cos\alpha_k(x_i; y_j, x_{i+1}) &\geqslant \frac{\Delta(x_i,y_j)}{|\nabla f|(x_i)} + C\delta \\ &\geqslant \left(1-\delta^2\right) + \frac{\Delta(x_i,y_j) - \Delta(x_i,x_{i+1})}{|\nabla f|(x_i)} + C\delta.\end{aligned}$$

Either $d(x_i,y_j) < \delta$ or $d(x_i,y_j) \geqslant \delta$ and

$$\Delta(x_i,y_j) - \Delta(x_i,x_{i+1}) \geqslant \delta - |\kappa| d(x_i,y_j).$$

From this last inequality it follows that for some constant C',

$$\frac{d(x_{i+1},y_j) - d(x_i,y_j)}{d(x_i,x_{i+1})} \leqslant C' d(x_i,y_j).$$

This is essentially a difference quotient; using an inductive argument is possible to show that pairs such that $|f(y_j) - f(x_i)| < \delta^2$ satisfy $d(x_i,y_j) \leqslant C''\delta$, as long as neither element is exceptional. But the exceptional points were chosen so that their contribution to the pairwise distances is small. These estimates allow one to show that any family of such sequences, with $\delta \to 0$, must converge to some curve. By the semicontinuity of the gradient, the resulting curve is a gradient curve. Uniqueness follows from the fact that a given gradient curve contains a sequence of the above type.

As we have observed before, the distance function is semiconcave. We will be mostly interested in gradient curves of the function d_p. Note that if $\gamma \in S_p$ then for any t, the segment of γ starting at t and going away from p is the gradient curve for d_p starting at $\gamma(t)$. Now let q be in the interior of γ. Then γ is a d_p-gradient curve, and there is a unique d_p-gradient curve σ starting at q. By uniqueness, γ and σ together must form a d_p-gradient curve. We have proved:

PROPOSITION 144. *Every minimal curve starting at $p \in X$ extends to a complete d_p-gradient curve in any neighborhood in which d_p has no critical points.*

We next state the result from [65], that representative angles have a similar monotonicity along gradient curves to that along geodesics.

DEFINITION 145. Given $p \in X$ and a curve c, let $\alpha_k^*(p; c(t_1), c(t_2))$ denote the angle at P of the triangle (P, C_1, C_2) in S_k such that $d(P, C_1) = d(p, c_1(t)), d(P, C_2) = d(p, c(t_2))$, and $d(C_1, C_2) = |t_1 - t_2|$. If c_1 and c_2 are curves such that $c_1(0) = c_2(0) = p$, let $\alpha_k^*(p; c_1(t_1), c_2(t_2))$ be the angle at p of the triangle (P, K_1, K_2) in S_k such that $d(P, K_1) = t_1, d(P, K_2) = t_2$, and $d(K_1, K_2) = d(c_1(t_1), c_2(t_2))$.

In other words, the above representatives use the parameter of the curve instead of distance, when any two points on the triple lie on one of the specified curves.

PROPOSITION 146. *Let γ_1, γ_2 be d_p-gradient curves starting at p. Then there exists a "proper" monotone reparameterization of γ_1, γ_2 on $[0, \infty)$ such that*

(1) *For any $q \in M$, $\alpha_k(p; q, \gamma_1(t))$ is nonincreasing in t.*
(2) *The function $\alpha_k^*(p; \gamma_1(s), \gamma_2(t))$ is nonincreasing in s, t.*

In other words, the same triangle comparisons hold for gradient curves as for geodesics if the parameter is used instead of the distance; the advantage is that gradient curves continue past the cut locus – until a critical point of d_p is met.

The parameterization of a d_p-gradient curve γ is obtained by beginning with a d_p-parameterization (cf. comments after Definition 140), then considering $\gamma \circ \rho^{-1}$, where $\rho(t) = \exp(I(t))$, where

$$I(t) = \int_0^t \frac{|(\nabla d_p)|^{-2} - 1}{\tau} \, d\tau.$$

We remark that this parameterization slows down rapidly as the curve approaches a point where $\nabla d_p = 0$; otherwise, the curve continues. In addition, $\rho(t)/t \to 1$ as $t \to 0$, so the initial derivative of the gradient curve is unchanged. We will need gradient curves tangent to more general directions. The following proposition can be proved using Propositions 146, 144, and Corollary 133.

PROPOSITION 147. *For any direction $\gamma \in \Sigma_p$ there exists a unique d_p-gradient curve $\tilde{\gamma}$ starting at p such that $\tilde{\gamma}'(0) = \gamma$. Furthermore, on any compact subset C of Σ_p there exists an $a > 0$ such that the gradient curves tangent to elements of C are uniformly Lipschitz on $[0, a)$.*

The latter notation $\tilde{\gamma}'(0) = \gamma$ means that shortest curves γ_t from p to $\tilde{\gamma}(t)$ converge in angle to γ. Thus when Σ_p is compact (which, as we will see, is true in the finite dimensional case), we can always construct a "gradient exponential map" that is locally surjective and distance decreasing (if T_p is given the metric of curvature $\geqslant k$ via the hyperbolic cone, given at the end of Section 4.5). Note that gradient curves do "bifurcate". For example, if there are two minimal curves from p to q then the two minimal curves extend as gradient curves; but by uniqueness, the two must coincide beginning at q.

8. Dimension

Throughout this section, X denotes an inner metric space of curvature $\geqslant k$ for some k.

8.1. *Spherical sets*

DEFINITION 148. A set of n points $\{a_1, \ldots, a_n\}$ in a metric space is called *semispherical* if $\det[\cos d(a_i, a_j)] > 0$. A set of $2n$ points $\{a_1, b_1, \ldots, a_n, b_n\}$ is called *spherical* if $\{a_1, \ldots, a_n\}$ is semispherical and $d(a_i, b_i) = \pi$ for all i.

In other words, a spherical set can be isometrically embedded in $S_1^n \subset \mathbb{R}^{n+1}$ as endpoints a_i of unit vectors forming a basis of $\mathbb{R}^{n+1}$, together with their antipodal points b_i.

THEOREM 149. *If X has curvature $\geqslant 1$ and contains a spherical set Σ of $2(n+1)$ points, then there is a subset S of X isometric to the unit sphere S^n such that $\Sigma \subset S$.*

The proof is a somewhat involved induction on n. We sketch some details. For $n = 0$, the proof is trivial. For $n = 1$ we essentially need to find a closed geodesic of length 2π through the four points such that each segment of length less than π is minimizing. First, we can join a_1 and b_1 by a minimal curve through a_2 as follows. Using Theorem 27 we choose points $x_i \to a_2$ joined to a_1, b_1 by minimal curves α_i, β_i, respectively. But then A1 implies (since we are comparing with the unit sphere) that $\alpha(\alpha_i, \beta_i) \to \pi$. It follows from the second monotonicity that the sequence has a uniform limit, which is the desired curve. We then complete the construction joining a_1, b_1 by a minimal curve through b_2. The fact that the resulting geodesic is an isometrically embedded circle follows from the observation that segments of it (considered as hinges of angle π) satisfy EA2 (see Proposition 35). The step from $n = 1$ to $n = 2$ is representative of the higher dimensional cases, and can be patterned after the proof of Proposition 35. We construct an embedded circle C containing a_1, a_2, b_1, b_2 using the inductive step. We then consider all minimal curves (and these can be constructed without local compactness, as above) from a_3, b_3 to C. Then EA2 can be verified for all hinges made of these curves; thus we have "filled in" a unit 2-sphere.

The standard hemisphere shows that not a single point can be removed from the spherical set without making the theorem false. This is in contrast to the situation for Riemannian manifolds, where Toponogov's Rigidity Theorem says that a Riemannian n-manifold of sectional curvature $\geqslant 1$ and diameter π must be isometric to S_1^n. It is interesting that smoothness is not a necessary assumption; the theorem can be proved using only geodesic completeness [74].

8.2. *Regularity*

We now use Theorems 27 and 149 to construct spheres in the space of directions at almost every point in X. The proof, in the present generality, is quite technical (see [76, Section 5]). We will attempt to make simplifications in the argument that do not obscure the general idea. For this argument (but not later!), the term "almost everywhere" will mean in a dense G_δ set. Note that if γ_{ab} is almost extendable beyond b, then by Proposition 55 there is a 0-sphere $\{\beta, -\beta\}$ in Σ_b. Thus by Theorem 27 there is a (unit) 0-sphere in the space of directions of almost every point (and one of the two directions in the 0-sphere is a geodesic direction). Now suppose there is a direction $\alpha \in \Sigma_b$ distinct from β and $-\beta$. Since S_b is dense in Σ_b, we can assume that α is a geodesic direction. It may be that α has no complement in Σ_b, so there may not be a spherical set of four elements in Σ_b. However, we can find such a spherical set in nearby spaces of directions as follows. Fix $t > 0$ small enough that both β and α restricted to $[0, t]$ are almost extendable past b (cf. Proposition 56). By Corollary 30 there is a dense G_δ set of points p joined to both $x = \beta(t)$ and $y = \alpha(t)$ by minimal curves almost extendable past p. But by Proposition 57, $\alpha(\gamma_{px}, \gamma_{py})$ must be close to $\alpha(\beta, \alpha)$ when p is close to b; hence we obtain a spherical set $\{a_1, b_1, a_2, b_2\} \subset \Sigma_p$, where $a_1 = \gamma_{px}, b_1 = -\gamma_{px}, a_2 = \gamma_{py}, b_2 = -\gamma_{py}$.

Now we are faced with the problem that Σ_p may not be an inner metric space, so we do not automatically get a 1-sphere in Σ_p from Theorem 149. However, from Proposition 53 we can find a minimal curve in Σ_p joining these two directions. In this way we can construct an inner metric space containing the spherical set as a convex subset, and complete the proof. By this argument, the given spherical set is contained in a sphere $S \subset \Sigma_p$. Either $S = \Sigma_p$, or there is another geodesic direction γ not in S. By Proposition 57 and the continuity of the determinant function, we can find a spherical set of $2(n+1)$ points (and hence an n-sphere) in Σ_p for points q near p.

In light of the above construction, the following definition makes sense:

DEFINITION 150. For any $p \in X$, we define the *regularity of* p to be $R(p) = \sup\{n\colon\ \Sigma_p$ contains a subset isometric to $S_1^{n-1}\}$. (We consider $S_1^{-1} = \emptyset$ and S_1^0 to be two points of distance π). We let $R^n(X) = \{p \in X\colon\ R(p) \geqslant n\}$. We define the *local dimension* of X to be $l\dim(X) = \sup\{n\colon\ R^n(X) \neq \emptyset\}$. A point p is said to be *regular* (of dimension n) if $\Sigma_p = S_1^{n-1}$.

The results of our construction can now be stated as follows:

THEOREM 151. *Let* $p \in X$, *and suppose that* $R(p) = n$. *If* p *is not regular, then* $R^{n+1}(X) \cap B(p, \delta)$ *contains a dense* G_δ *for all small* $\delta > 0$.

8.3. *Equivalence of Hausdorff and covering dimension*

Our next task is to relate local dimension to two other standard notions of dimension: Hausdorff dimension and covering dimension. See [47] for the definitions of these two dimensions. We restrict ourselves to the following comments: We will denote Hausdorff dimension by $h\dim$ and covering dimension by dim. One can define the n-dimensional Hausdorff measure υ_n of a metric space; υ_n is decreased by distance decreasing functions. Then $h\dim(X) = \sup\{m\colon\ \upsilon_m(X) > 0\}$. It follows that distance nonincreasing maps do not increase Hausdorff dimension. Also, $h\dim$ depends on the metric, not just the topology, of X. The following lemma is a consequence of the definitions:

LEMMA 152. *If Y is a metric space then*
(1) *Scaling the metric of Y does not change its Hausdorff dimension.*
(2) $h\dim Y = n$ *if and only if* $h\dim cY = n+1$.
(3) *If Y is the Gromov–Hausdorff limit of metric spaces Y_i,* $h\dim Y \leqslant \liminf h\dim Y_i$.

On the other hand, dim is defined for topological spaces, and is a topological invariant. Any topological space Y containing an open subset homeomorphic to an open subspace of $\mathbb{R}^n$ has $\dim Y \geqslant n$. For any metric space Y, $\dim Y \leqslant h\dim Y$. Indeed, $\dim Y = \inf\{h\dim(Y, d)\}$, where the infimum is over all metrics d compatible with the given topology. Every n-dimensional Riemannian manifold has Hausdorff dimension n with respect to the distance induced by the Riemannian metric, and each finite metric ball has finite n-dimensional Hausdorff measure.

If $l\dim(X) = n$ then by Theorem 151 we can find a regular point p, i.e., $T_p = \mathbb{R}^n$. For any k, we can put on $B(0, \pi/\sqrt{k}) \subset \mathbb{R}^n$ a Riemannian metric of constant sectional curvature k, such that the radial lines are still geodesics, using one of the cone constructions at the end of Section 4.5. Equipped with such a metric, it is immediate from Theorem 43 and A2 that $\exp_p$ is distance decreasing. Since it is distance decreasing, we can extend it continuously to a mapping $\overline{\exp_p}$ defined on the closure C of the domain of $\exp_p$. If we let C_r be the compact set $C \cap \overline{B}(0, r)$ then $\overline{\exp_p}(C_r)$ is compact, and equals $\overline{B}(p, r)$. It follows that X is locally compact, and therefore, by Theorem 8 $\exp_p$ is surjective. Since the domain of $\exp_p$ is contained in the space T_p, which has Hausdorff dimension n, we see that $h\dim(X) \leqslant n$. More strongly, we have proved:

PROPOSITION 153. *If $l\dim X = n$, then X is locally compact, and every compact subset C of X has $\upsilon_n(C) < \infty$.*

To show that regular points are manifold points, we need to use the *distance coordinates*, which were discovered by Berestovskii [6].

DEFINITION 154. Let $x_1, \dots, x_n$ be points in X. We define the *distance coordinates* associated with $\{x_1, \dots, x_n\}$ to be the (obviously continuous) mapping $\Delta : X \to \mathbb{R}^n$ given by $q \mapsto (d(q, x_1), \dots, d(q, x_n))$.

We are interested in determining when the distance coordinates are indeed a coordinate system near some points (i.e., a homeomorphism onto its image). In the case Berestovskii considered, he already knew that X was a manifold by other means; having shown that Δ was 1–1 was sufficient, by invariance of domain, to conclude that Δ was a homeomorphism. In the present case, our argument for surjectivity is based on that from [19, Theorem 5.4].

Suppose that $l\dim(X) = n$, and let $p \in X$ be such that $\Sigma_p = S_1^{n-1}$. Then for any small $\delta > 0$ and all $1 \leqslant i \leqslant n$ we can find unit geodesics $\gamma_i = \gamma_{px_i}$ and $\beta_i = \gamma_{py_i}$ of fixed length $r > 0$ such that $|\alpha(\gamma_i, \gamma_j) - \pi/2| < \delta$ if $i \neq j$ and $|\alpha(\gamma_i, \beta_i) - \pi| < \delta$. Furthermore, we suppose that each γ_i, β_i is almost extendable past p and is the unique minimal curve between its endpoints (cf. Proposition 56). This special kind of "almost" spherical set was called an (n, δ)-strainer in [18], and an (n, δ)-explosion in [19]. We claim that the distance coordinates Δ associated with $\{x_1, \ldots, x_n\}$ are a homeomorphism of a small open set about p onto its image in $\mathbb{R}^n$. First, suppose Δ is not locally 1–1. Then there exist $a_j, b_j \to p$ such that $d(a_j, x_i) = d(b_j, x_i)$ for all i, j. Let ψ_{ji} be minimal from a_j to x_i, ξ_{ij} be minimal from a_j to y_i, and η_j be minimal from a_j to b_j. Then on the one hand it is immediate from Proposition 57 that if a_j is near p, the curves ψ_{ij} are a semispherical set in Σ_{a_j} (i.e., they are nearly orthogonal). By Proposition 63, $\lim \alpha(\psi_{ij}, \eta_j) = \pi/2$, so ψ_{ij} together with η_j form a semispherical set Σ of $n+1$ elements. If the elements of Σ all had complements (and they might not), we would have a contradiction to Theorem 151. However, note that $\alpha(\psi_{ij}, \xi_{ij})$ can be made arbitrarily close to $\pi - \delta$. Then using a technical modification of Proposition 57 it is possible to prove that we can find, in the space of directions at points near a_j, a spherical set of $2(n+1)$ elements, which contradicts Theorem 151. Thus Δ is locally 1–1.

To see why Δ is locally surjective, let x be close to p and suppose that $y = (y_1, \ldots, y_n)$ is close to $\Delta(x)$. We need to find a point x' close to x such that $\phi(x') = y$. First, $|d(x_i, x) - y_i|$ is maximal for some i. Join x to x_j by a minimal curve κ_j and x to y_j by a minimal curve λ_j. If x is close enough to p then $|\alpha(\kappa_i, \kappa_j) - \pi/2| < 2\delta$ if $i \neq j$ and $|\alpha(\kappa_i, \lambda_i) - \pi| < 2\delta$. Now we move along either κ_i or λ_i until we have reached a point z_1 such that $d(z_1, x_i) = y_i$ (this can be done by the Intermediate Value Theorem). In doing so, since we are moving along a curve that is almost orthogonal to κ_j for $j \neq i$, we see that $d(z_1, x_j)$ is only slightly changed from $d(x, x_j)$. We continue in this way, constructing a Cauchy sequence $\{z_i\}$ which converges to the desired point x'. We have now proved:

PROPOSITION 155. *If $l\dim(X) = n$ and $\Sigma_p = S_1^{n-1}$ then a small neighborhood of p is homeomorphic to an open subset of $\mathbb{R}^n$, via the distance coordinates of appropriately chosen points near p.*

The final result of this section shows that all notions of dimension are equivalent for spaces of curvature $\geqslant k$. The key remaining part of the proof, that finite topological dimension implies finite Hausdorff dimension, is from [65].

THEOREM 156. *Let X be a space of curvature $\geqslant k$. Then for any n, the following are equivalent*:

(1) $l\dim X = n$,

(2) $h\dim X = n$,

(3) $\dim X = n$,

(4) *for some point $p \in X$, $\Sigma_p = S_1^{n-1}$*,

(5) *there exists a dense G_δ set of points $p \in X$ such that $\Sigma_p = S_1^{n-1}$*,

(6) *X has a dense set of n-manifold points.*

Furthermore, when the above conditions hold for some n, X is locally compact.

So far we know the following: From Theorem 151 we get (1) $\Rightarrow$ (4). From the proof of Proposition 153 we get

$$(1) \Rightarrow (4) \Rightarrow h\dim X \leqslant n. \tag{7}$$

From Proposition 155 we get

$$(1) \Rightarrow (4) \Rightarrow \dim X \geqslant n. \tag{8}$$

Putting (7) and (8) together we get (4) $\Rightarrow$ (2) and (3). Now if $h\dim X < n$ then $\dim X < n$ so (8) implies $l\dim X \neq n$. If $h\dim X > n$ then we reach the same conclusion from (7) and we have proved (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (4). Now (5) $\Rightarrow$ (4) is obvious. To prove the opposite implication, note first that by Proposition 153 and (1), X is locally compact, and hence separable. Choose a countable dense subset Y of X, and let $K = \bigcap_{y\in Y} J_y$. Certainly K is a dense G_δ. An easy modification of the proof of Theorem 151 shows that $\Sigma_p = S_1^{n-1}$ for all $p \in K$. (5) $\Rightarrow$ (6) is an immediate consequence of Proposition 155. We need only prove two more implications: (3) $\Rightarrow$ (1) and (6) $\Rightarrow$ (1) .

Both are proved in the same way. As was pointed out in [65], the dimension argument of Proposition 2.10, [72] can now be carried out by using the "gradient exponential map" rather than the usual exponential map. Suppose $l\dim X \geqslant m$, and let $p \in X$ be such that Σ_p contains a copy S of S_1^{m-1}. Then by Proposition 147, there exists a number $a > 0$ and, in each direction γ of S, a corresponding proper gradient curve $g_\gamma : [0, a) \to X$. Note that we do not know if $S_p \cap S$ is dense in S. However, we can find finite subsets Y_i of S_p which are better and better approximations of S; the union $Y := \bigcup Y_i$ is precompact, and $S \subset \overline{Y}$. Now choose a finite ε-net $E_\varepsilon = \{\gamma_j\}$ of elements of Y. Since E is finite, the cut radius C_p has a positive lower bound R_ε on E_ε. If ε is small, then Proposition 146 implies that gradient curves tangent to vectors in $\overline{Y}$, and hence in S, do not meet on $(0, R_\varepsilon)$ unless their angle is small. In other words, if $\phi_\rho : B(0, a) \subset \mathbb{R}^m \to X$ is defined by $\phi_\rho(tv) = g_v(\rho t)$, then the inverse images of points are uniformly small for ρ small. By [47, IV.5.A], this means $\dim X \geqslant \dim B(0, a) = m$. In fact, the dimension of any small ball near p has dimension $\geqslant m$. So if $\dim X = n$ then $n \geqslant l\dim X = h\dim X \geqslant \dim X = n$, and we have proved (3) $\Rightarrow$ (1). We can apply the same argument to an n-manifold point to get (6) $\Rightarrow$ (1).

9. Alexandrov spaces

Finite dimensional spaces of curvature $\geqslant k$ have become known simply as "Alexandrov Spaces". From the last section we know that "finite dimensional" means that the topological, Hausdorff, or local dimension is finite. We know that such a space is locally compact, has a dense G_δ subset of points in which the space of directions is a sphere of one dimension lower, and has an open dense subset of manifold points. In this chapter we discuss further the structure of Alexandrov spaces. Throughout this section, X denotes an Alexandrov space of dimension n.

9.1. *Basic properties*

THEOREM 157. *If $n \geqslant 2$ then for any $p \in X$, Σ_p is an $(n-1)$-dimensional Alexandrov space of curvature $\geqslant 1$.*

To prove this theorem, first observe that the space of directions at every point is compact. This follows from the fact that the Σ_p is a sphere at a dense set of points and the semi-continuity of the angle (Lemma 36). In fact, if Σ_p were not compact at some point p, we could find a sequence $\{\gamma_i\}$ in S_p such that $\alpha(\gamma_i, \gamma_j) > \varepsilon$ for all i, j. Let p_i be the endpoint of γ_i different from p, $y_j \to p$ be such that Σ_{y_j} is isometric to S_1^{n-1}, and γ_{ij} be minimal from y_j to p_i. Then by Lemma 36 for any j, k, $\alpha(\gamma_{ij}, \gamma_{ik}) > \varepsilon/2$ for large enough i. In other words, we can construct in S_1^{n-1} a finite sequence of points with arbitrarily many elements having pairwise distance $> \varepsilon/2$. This is impossible, since S_1^{n-1} is compact. We now know that Σ_p is compact for all $p \in X$. The proof of Theorem 157 is now finished by Proposition 87, Lemma 152 and the following proposition.

PROPOSITION 158. *For any* $p \in X$, $(T_p, 0)$ *is the pointed Gromov–Hausdorff limit of the scaled metric spaces* (X_n, p), *where* X_n *is* X *with the scaled metric* nd, *as* $n \to \infty$.

The proof of the above proposition uses Proposition 74 and the compactness of Σ_p, proved above. In fact, this makes $\overline{B}(0, r) \subset T_p$ compact for any $r > 0$. By moving it slightly, we can assume that an ε-net $\mathcal{N}$ in $\overline{B}(0, r)$ is of the form $s_i\gamma_i$, where each $\gamma_i \in S_p$ and $s_i \leqslant r$, and $i \in 1, \dots, m$. Now for any large k,

$$\left(d\left(\gamma_i\left(\frac{s_i}{k}\right), \gamma_j\left(\frac{s_j}{k}\right)\right) - \sqrt{\left(\frac{s_i}{k}\right)^2 + \left(\frac{s_j}{k}\right)^2 - 2\left(\frac{s_i}{k}\right)\left(\frac{s_j}{k}\right)\cos\alpha(\gamma_i, \gamma_j)}\right)/rk$$

is uniformly small (cf. Lemma 33), and so in the scaled metric, $\{\gamma_i(s_i/k)\}$ is Lipschitz close to $\mathcal{N}$.

Theorem 157 allows the useful technique of making definitions and proving theorems by induction on the dimension of the space.

We can now improve on the "first variation formula", Proposition 61, in the finite dimensional case. Note that for an Alexandrov space A'_p (Definition 122) is compact. Here we are using the notation of Section 7.4.

PROPOSITION 159. *Let* $p \neq q$ *in* X. *Then for any* $\gamma \in \Sigma_q$, $d(d_p)_q(\gamma) = -\cos\alpha(\gamma, q')$.

In particular, if γ is a geodesic direction then, by definition, the (right-hand) derivative of $d_p(\gamma(t))$ at $t = 0$ is equal to $-\cos\min\{\alpha(\gamma, \beta)$: β joins p and $q\}$. Proposition 61 implies that, for $\gamma \in S_p$, $d(d_p)_q(\gamma) \leqslant -\cos\alpha(\gamma, q')$. On the other hand, if we choose any geodesics β_t from p to $\gamma(t)$, these, by the compactness of the space of directions, must converge in angle to some minimal curve joining p and q. Applying Proposition 52 and Proposition 61 gets the opposite inequality.

The last result in this section concerns the measure of regular points, and was proved in [60]. We defined *regular point* in Definition 150.

DEFINITION 160. A point which is not regular is called *singular*. The set of singular points in X is denoted by $\mathcal{S}_X$.

This theorem is a strengthening of the fact nonregular, or *singular*, points are nowhere dense.

THEOREM 161. *The set $\mathcal{S}$ of singular points in X has $h\dim(\mathcal{S}) \leqslant n-1$.*

9.2. ε-*Open maps*

Here we recount some definitions and results from [19], Sections 5 and 11, about ε-open maps. Note that if Y, Z are metric spaces and $U \subset Y$ is open, a mapping $\phi : U \to Z$ is open if and only if for every $x \in U$ there exists a $\rho > 0$ such that if $y \in B(\phi(x), \rho)$ then there exists an $x' \in U$ such that $\phi(x') = y$. The next definition clearly strengthens the notion of open map.

DEFINITION 162. Let Y, Z be metric spaces, $U \subset Y$ be open and $x \in U$. We define $\rho(x, U) := \sup\{r\colon\ B(x, r) \subset U\}$. A mapping $\phi : U \to Z$ is called *ε-open* if for every $y \in B(\phi(x), \varepsilon\rho(x, U))$, there exists an $x' \in X$ such that $\phi(x') = y$ and $d(x, x') \leqslant \frac{1}{\varepsilon} d(\phi(x), y) < \rho(x, U)$.

Note that if Z is an inner metric space and Y is locally compact then we can replace $B(\phi(x), \varepsilon\rho(x, U))$ with $\overline{B}(\phi(x), \varepsilon\rho(x, U))$ in the above definition (cf. comments prior to Theorem 8). The proofs of the next lemmas are obvious:

LEMMA 163. *Let W, Y be metric spaces, $U \subset W$ be open and $\phi : U \to Y$ be ε-open. Then for all $x \in U$, $B(\phi(x), \varepsilon\rho(x, U)) \subset \phi(B(x, \rho(x, U)))$.*

LEMMA 164. *Let W, Y be metric spaces, $U \subset W$ be open and $\phi : U \to Y$ be* 1–1, *continuous, and ε-open. Then for all $x \in W$, the mapping $\phi^{-1} : B(\phi(x), \varepsilon\rho(p, U)) \to U$ is a Lipschitz homeomorphism (onto its image) with Lipschitz constant $1/\varepsilon$.*

DEFINITION 165. A Lipschitz function $f : U \subset X \to \mathbb{R}$ is called *directionally differentiable* if for each $p \in U$ there exists a continuous function $df_p : \Sigma_p \to \mathbb{R}$ such that for every $\gamma \in S_p$, $(f \circ \gamma)^{+}_{t=0}$ exists and is equal to $df_p(\gamma)$.

Note that df_p can be extended "linearly" to T_p. From Proposition 128 we know that semiconcave functions are directionally differentiable. As in the case of semiconcave functions we will omit the subscript "p" when no confusion will result. We are now in a position to extend Proposition 155.

PROPOSITION 166. *Let $f_1, \ldots, f_m : U \subset X \to \mathbb{R}$ be a collection of directionally differentiable functions. Suppose that there exist δ, ε with $0 < n\delta < \varepsilon$ such that for any $p \in U$ there exist directions $\gamma_i^{+}, \gamma_i^{-} \in \Sigma_p$ for $i = 1, \ldots, m$ with the following property: for $i \neq j \in \{1, \ldots, m\}$, $df_i(\gamma_i^{+}) > \varepsilon$, $df_i(\gamma_i^{-}) < -\varepsilon$, and $|df_i(\gamma_j^{\pm})| < \delta$. Then the map $f := (f_1, \ldots, f_m)$ is σ-open, where $\sigma := \frac{\varepsilon - n\delta}{2}$, with respect to the norm $\|v\| = \sum |v_i|$ in $\mathbb{R}^m$.*

We remind the reader that in the case of Proposition 155 we were dealing with distance functions, so we were able to use an almost spherical set to prove the surjectivity of the mapping. In the present case the idea is similar: the directions $\gamma_i^{\pm}$ and the values of df_i in these directions force the functions to map onto an open region in $\mathbb{R}^m$. To prove it, for $x \in X$, let $B = \overline{B}(f(x), \sigma\rho(x, U))$ and let $A \subset B$ be the set of all $y \in B$ such that there exists an x' such that $d(x, x') \leqslant \frac{1}{\sigma} d(f(x), y)$ and $f(x') = y$. Then A is nonempty ($x \in A$) and closed. If the proposition were false, then for some x there would exist a $z \in B \backslash A$. Let y be the element of A closest to z and x' be such that $f(x') = y$ and $d(x, x') \leqslant \frac{1}{\sigma} d(f(x), y)$. Suppose, for definiteness, that $z^i < y^i$. Now by the assumptions of the proposition we can find a geodesic direction γ at x' near some particular γ_i^+; i.e., so that $df_i(\gamma) > \varepsilon$ and $|df_j(\gamma)| < \delta$. Now by moving a short way along the geodesic γ we get a point x'' whose image $y'' = f(x'')$ contradicts the choice of y as the closest point in A to z.

9.3. *Stratification and local cone structure*

Two fundamental theorems about the topological structure of Alexandrov spaces are due to Perelman [62,63]. We will sketch the arguments in subsequent sections.

THEOREM 167. *A small open sphere at any point in* X *is homeomorphic to the cone on its boundary.*

THEOREM 168. X *possesses a canonical stratification with strata homeomorphic to topological manifolds.*

These two theorems are proved using a kind of generalized Morse theory for the distance function. Morse theory for the distance function was discovered in the now famous paper of Grove and Shiohama [44], see also [35]. The first theorem of Morse theory is that if M is a Riemannian manifold and $f : M \to \mathbb{R}$ is smooth, proper (i.e., preimages of compact sets are compact) and without critical points (see Definition 121 and the comments before Proposition 142) in the region $f^{-1}([a, b])$ then $M^a := f^{-1}((-\infty, a])$ is diffeomorphic to $M^b := f^{-1}((-\infty, b])$, and the former is a deformation retract of the latter. The idea of the proof is to retract M^b to M^a along the gradient curves of f. Another way to state this result is that f is a trivial bundle map on $f^{-1}([a, b])$:

DEFINITION 169. If Y is a topological space, a map $f : Y \to \mathbb{R}$ is a *(locally trivial) bundle map* if for each $p \in Y$, there is a neighborhood U of p homeomorphic to a product $f(U) \times C$ such that f restricted to U is the projection onto $f(U)$. A bundle map f is called *trivial* if we can take $U = Y$.

The problem with applying Morse theory to the distance function, even for Riemannian manifolds, lies in the fact that the distance function d_p is not smooth at points in the cut locus of p. In an Alexandrov space, the problem is manifested in the behavior of the gradient curves of the distance function which, as we observed at the end of Section 7.4, may bifurcate.

9.4. *DER functions*

In this subsection we denote by Σ an Alexandrov space of curvature $\geqslant 1$. In order to extend the first theorem of classical Morse theory to Alexandrov spaces, Perelman introduces in [63] a class of functions, called *admissible maps*, which include distance functions. Admissible maps satisfy a slightly stronger condition than κ-convexity (Definition 124), and have differentials on Σ_p which can be written as finite formal sums

$$\sum_i -a_i \cos d_{A_i}, \tag{9}$$

where as usual d_{A_i} denotes the distance from the compact set A_i, and a_i are non-negative constants such that $\sum a_i \leqslant 1$.

DEFINITION 170. Expressions of the form (9) on Σ are called functions of class DER or simply DER functions. When no confusion will occur, we will not explicitly mention the space Σ. Note that a given function may be expressed as a DER function in more than one way.

We list several properties of DER functions, which can be proved by direct computation:

PROPOSITION 171. *If $f = \sum_i -a_i \cos d_{A_i}$ is a DER function then for all $x, y, z \in \Sigma$ such that y is between x and z,*
(1) $|f(x)| \leqslant 1$,
(2) $|f(x) - f(y)| \leqslant d(x, y)$,
(3) $\sin d(x, y) f(y) \geqslant \sin d(x, y) f(z) + \sin d(y, z) f(x)$,
(4) df_x *is the DER function* $\sum_i a_i \sin d(x, A_i) \cos d_{A_i'}$.

The last part of the proposition allows one to prove statements about this class of functions by induction.

It is useful to define the following "scalar product" on DER. Given $f = \sum_i -a_i \cos d_{A_i}$ and $g = \sum_j -b_j \cos d_{Bj}$, let $\langle f, g\rangle = \sum_{i,j} a_i b_j \cos d(A_i, B_j)$. The scalar product is bilinear (although one should keep in mind that DER is not closed under addition and positive scalar multiplication due to the restriction $\sum a_i \leqslant 1$). If one defines the *characteristic functions* $\chi_p := -\cos d_p$ and $\chi_A := -\cos d_A$, for any $p \in \Sigma$ or $A \subset \Sigma$, one immediately has

$$\langle f, \chi_p\rangle = -f(p) \quad \text{and} \quad \langle f, \chi_A\rangle \geqslant -f(q) \quad \text{for all } q \in A. \tag{10}$$

We now can state the following lemma (in it, and in the rest of this section, we consider only the differential restricted to Σ_p).

LEMMA 172. *If $f, g \in DER$ then $\langle df, dg\rangle \leqslant \langle f, g\rangle - f(p)g(p)$.*

The lemma is proved by applying bilinearity and the following lemma, which is an immediate consequence of the definitions, formula (1), and A2.

LEMMA 173. *For any* $A, B \subset \Sigma$,

$$\cos d(A, B) \geqslant \cos d(p, A) \cos d(p, B) + \sin d(p, A) \sin d(p, A) \cos d(A', B').$$

Questions about DER functions can often be reduced to ones about characteristic functions. We make some observations that will facilitate this reduction. First, if $f = \sum_i -a_i \cos d_{A_i}$ and $g = \sum_j -b_j \cos d_{B_j}$, then picking $p_i \in A_i$, we immediately have that $\langle \sum_i -a_i \chi_{p_i}, g\rangle \leqslant \langle f, g\rangle$. More strongly,

LEMMA 174. *Given any DER function on* Σ *there exists a function* $a\chi_p$ *such that* $\langle f, g\rangle \geqslant \langle a\chi_p, g\rangle$ *for any DER function* g *on* Σ.

The observation before the lemma allows one to reduce the proof to the case $f = a_1\chi_{p_1} + a_2\chi_{p_2}$, where $p_1 \neq p_2$. If $d(p_1, p_2) = \pi$ then assuming, say, $a_1 \leqslant a_2$, one can pick $p = p_1$ and $a = a_1 - a_2$. Otherwise, Proposition 171(3) implies that the point p on $\gamma_{p_1 p_2}$ and a such that $a/\sin d(p_1, p_2) = a_1/\sin d(p, p_2) = a_2/\sin d(p, p_1)$ will work. From the above lemma we can now finish the properties of the scalar product:

PROPOSITION 175. *The scalar product* $\langle *, * \rangle$ *on the space of DER functions on* Σ *is a symmetric, bilinear operator such that*

$$\langle f, f\rangle \geqslant \Bigl(\inf_{g \in DER} \langle f, g\rangle\Bigr)^2 \geqslant 0$$

for every DER function f.

The main result about DER functions is the following:

PROPOSITION 176. *Let* Σ *be an* n*-dimensional Alexandrov space of curvature* $\geqslant 1$, $n \geqslant 0$. *Let* f_i *be DER functions on* Σ *for* $i = 0, \dots, k+1$ *for some* $k \geqslant 0$. *If* $\varepsilon := \min_{0 \leqslant i \neq j \leqslant k+1}\{-\langle f_i, f_j\rangle\} > 0$ *then*

(1) *there exists a* $q \in \Sigma$ *such that* $f_i(q) \geqslant \varepsilon$ *for all* $0 \leqslant i \leqslant k$,
(2) $k \leqslant n$, *and*
(3) *there exists a* $p \in \Sigma$ *such that* $f_i(p) = 0$ *for* $1 \leqslant i \leqslant k$, $f_0(p) \geqslant \varepsilon$ *and* $f_{k+1}(p) \leqslant -\varepsilon$.

To prove the first part note that by Lemma 174 and formula (10) we can find a function $a\chi_q$ such that $f_j(q) = -\langle a\chi_q, f_j\rangle \geqslant -\langle f_{k+1}, f_j\rangle \geqslant \varepsilon$. The second part can be proved by induction, noting that the derivatives of functions satisfying the hypotheses of the proposition again satisfy the hypotheses. The point p can be found as a point where f_0 attains its maximum value among all points where $f_i \geqslant 0$ for $i = 0, \dots, k$. If one of the functions f_i, $i = 1, \dots, k$, doesn't vanish at p, Lemma 172 implies that the first part of the present proposition (proved above) can be applied to the differentials $\{df_i\}$ at p, $i = 0, \dots, k$. We then find a geodesic direction γ such that $df_i(\gamma) > 0$ for all $i = 0, \dots, k$. Moving along γ then contradicts the maximality of $f_0(p)$. The condition on f_{k+1} can be proved similarly.

9.5. *Admissible functions*

DEFINITION 177. A directionally differentiable map $f : U \subset X \to \mathbb{R}$ is called *admissible* if df_p is a DER function on Σ_p for all p, and there exists a $\lambda \in \mathbb{R}$ such that for every $p, q \in U$,

$$f(q) \leqslant f(p) - \langle df, \chi_{q'} \rangle_p d(p,q) - \lambda d(p,q)^2. \tag{11}$$

Note that κ-convexity implies, by Lemma 112(8), $f(q) \leqslant f(p) + df(\gamma)d(p,q) + |\kappa| d(p,q)^2$ for every $\gamma \in q'$. Since $-\langle df, \chi_{q'} \rangle \leqslant df(\gamma)$ by Lemma 10 we see that Definition 177 strengthens κ-convexity. The main examples of admissible functions are the distance functions. If $A \subset X$, an easy generalization of Proposition 159 is that, given $p \in X$ and $\gamma \in \Sigma_p$ if $f := d_A$ then, $df_p = \chi_{A'_p}$. In other words, by formula (10), $-\langle df, \chi_{q'} \rangle_p = -\cos\alpha(A'_p, q')$. Since $\alpha(A'_p, q') = \min\{\alpha(\gamma_{pq}, A'_p)\}$, formula (11) reduces to κ-convexity. We have proved:

PROPOSITION 178. *For any closed $A \subset X$, the function d_A is admissible, and $d(d_A)_p = \chi_{A'_p}$.*

DEFINITION 179. A map $g : X \to \mathbb{R}^k$ is called *admissible* in U if $g = G \circ (f_1, \dots, f_k)$ where each f_i is admissible in the sense of Definition 177 and G is a Lipschitz homeomorphism of open sets of $\mathbb{R}^k$. A point $p \in U$ is called ε-regular for g if $f := (f_1, \dots, f_k)$ satisfies the following properties:

(1) If q is sufficiently close to p and $i \neq j$ then $\langle f'_i, f'_j \rangle_q < -\varepsilon$, and
(2) $\Sigma^\varepsilon_p := \{\gamma \in \Sigma_p : df_i(\gamma) > \varepsilon \text{ for all } 1 \leqslant i \leqslant k\} \neq \emptyset$.

A point p is called a *regular point of* g if p is ε-regular for some $\varepsilon > 0$. If $w \in \mathbb{R}^k$ has the property that every $f^{-1}(w)$ is nonempty and contains only regular points then x is called a *regular value* of f.

REMARK 180. We will say that every mapping $f : X \to \mathbb{R}^0$ is admissible without critical points. Note that if $k = 1$ then a point p is a regular point of g if and only if it is not a critical point of the admissible map f_1. This follows from Corollary 134.

Note that, since an admissible function is κ-convex, condition (2) above implies that $\Sigma^\varepsilon_q \neq \emptyset$ for all q sufficiently close to p. It follows that the set of regular points of an admissible function is open. However, the set of regular points of g *may be empty*. The functions $(df_i)_q$ and χ_γ, where $\gamma \in \Sigma^\varepsilon_q$, satisfy the conditions of Lemma 176, so we can find directions $\gamma_i^\pm \in \Sigma_q$ such that $df_i(\gamma_i^\pm) > \varepsilon$ and $df_i(\gamma_j^\pm) = 0$ whenever $i \neq j$. Applying Proposition 166 we now get:

PROPOSITION 181. *If p is an ε-regular point of an admissible map g then g is ε-open in a neighborhood of p.*

DEFINITION 182. A function $g: X \to \mathbb{R}^k$ is called *incomplementable* at a regular point p if p is not a regular point of any admissible map $(g, g_{k+1}): U \to \mathbb{R}^{k+1}$, where U is a neighborhood of p.

PROPOSITION 183. *Suppose $g: X \to \mathbb{R}^k$, $k \leqslant n$, is an admissible map in a neighborhood of a regular point p, which is incomplementable at a regular point p. Then there exists a continuous nonpositive function $g_{k+1}: X \to \mathbb{R}$ which vanishes at p, and a neighborhood U of p, such that $k := (g, g_{k+1}): X \to \mathbb{R}^{k+1}$ is admissible in U and*

(1) *for small enough $\rho > 0$ the set $K_\rho := U \cap g_{k+1}^{-1}([0, -2\rho]) \cap g^{-1}(\overline{B}(g(p), \rho))$ is compact,*
(2) *if $w \in g(K_\rho)$, then $K_\rho \cap g_{k+1}^{-1}(0) \cap g^{-1}(w)$ is a point, and*
(3) *every point of $K_\rho \backslash g_{k+1}^{-1}(0)$ is a regular point of k.*

The proof of this proposition may be found in [63].

REMARK 184. Note that $K_\rho \backslash g_{k+1}^{-1}(0) \neq \emptyset$ if $k < n$. Otherwise, by condition (2) above, $g: K_\rho \to \mathbb{R}^k$ would be 1–1, which is impossible given the difference in dimension.

9.6. *Topological arguments*

This section finishes the proofs of Theorems 168 and 167. In this section we consider the open cone on the empty set $\emptyset$ to be a point. A *conical neighborhood* of a point p in a topological space Y is an open set U containing p which is (pointed) isomorphic to the open cone cZ on a topological space Z; i.e., U is homeomorphic to cZ via a homeomorphism h such that $h(p)$ is the apex 0 of the cone. Conical neighborhoods are unique up to pointed homeomorphism [48].

DEFINITION 185. A (-1)-dimensional *MCS-space* is defined to be the empty set. If $n \geqslant 0$ then an n-dimensional *MCS-space* is a metric space, each of whose points has a conical neighborhood, pointed homeomorphic to the open cone over a compact $(n-1)$-dimensional MCS-space.

For example, a 0-dimensional MCS-space is discrete and a 1-dimensional MCS-space is a graph. Clearly the cone or suspension of an n-dimensional MCS-space is an $(n+1)$-dimensional MCS-space. MCS-spaces are stratified by manifolds: The m-dimensional stratum Z_m consists of those points whose conical neighborhood splits as $\mathbb{R}^m \times cZ$, where Z is a compact MCS-space, and m is the largest number so that such a splitting exists. To see why Z_m is a manifold, note that the mapping $\phi_t: \mathbb{R}^m \times cZ \to \mathbb{R}^m \times cZ$ given by $\phi_t(s, x) = (s + t, x)$ is a homeomorphism. Therefore any two points $(s, 0)$ and $(t, 0)$ have conical neighborhoods isomorphic to $\mathbb{R}^m \times cZ$, and points not in the slice of 0 have conical neighborhoods isomorphic to $\mathbb{R}^{m+1} \times Z$. It follows that if $x \in Z_m$ then the points on the $\mathbb{R}^m$ slice of the conical neighborhood of x are also in Z_m. The remaining points near x are in Z_k for $k > m$. In particular, MCS-spaces are locally connected WCS-sets in the sense of [83, Definition 5.1].

THEOREM 186. *Let $g : X \to \mathbb{R}^k$ be an admissible map, $k \leqslant n$. Then*
(1) *g is a trivial bundle map in some neighborhood of any regular point p,*
(2) *if $w \in \mathbb{R}^k$ is a regular value for g then $g^{-1}(w)$ is an $(n-k)$-dimensional MCS-space,*
(3) *if g is proper and without critical points in U then f is a bundle map on U.*

Recall that a map is *proper* if the preimage of every compact set is compact; a closed map is proper if the preimages of points are compact. The proof of the theorem is by reverse induction in k; we denote the above three statements by $1(k)$, $2(k)$, and $3(k)$. For any k, $1(k)$ and $2(k)$, together with the facts that MCS-spaces are locally connected WCS-spaces (so [83, Theorem 5.4], can be applied) and Lipschitz maps are closed, allow Corollary 6.14 of [83] to be applied, and $3(k)$ follows. Statements $1(n)$ and $2(n)$ follow from Proposition 181. We suppose that conditions $1(k+1)$, $2(k+1)$, and $3(k+1)$ hold for some $k+1 \leqslant n$. If g is complementable then there is no problem. If g is incomplementable at a regular point p, consider the map g_{k+1} and the compact set K_σ given by Proposition 183. Since every point of K_σ is a regular point of (g, g_{k+1}) applying part $3(k+1)$ to this map we get a mapping $\phi : M \times \overline{B}(g(p), \rho) \times [-2\rho, 0) \to K_\rho - g_{k+1}^{-1}(0)$, where $M := K_\rho \cap g_{k+1}^{-1}(-\rho) \cap g^{-1}(g(p))$. But since $K_\rho \cap g_{k+1}^{-1}(0) \cap g^{-1}(v)$ is a single point for all $v \in g(K_\rho)$, we can extend this trivialization to $cM \times I \to K_\rho$. This implies $1(k)$ and $2(k)$.

Theorem 168 now follows from the statement $2(0)$. Since d_p has no critical points near p (cf. Corollary 133).

9.7. *Extremal sets*

In this section we mention some results from [66] about extremal sets. Extremal sets were introduced to further refine the stratification of Alexandrov spaces.

DEFINITION 187. A closed subset F of X is called *extremal* if for every $q \in X$, if d_q has a local minimum at p when restricted to F, then p is a critical point of d_q on X. If X has positive curvature (scaled to curvature $\geqslant 1$) we assume that $\overline{B}(F, \pi/2) = X$.

REMARK 188. It is possible in many cases to avoid the latter assumption (and simplify the definition) by working with the tangent cone instead of the space of directions (cf. [69]).

DEFINITION 189. Let $p \in X$. We define $K(p) = \{q \in X\colon q$ has a conical neighborhood isomorphic to a conical neighborhood of $p\}$.

PROPOSITION 190. *Let F be a closed subset of X such that if $p \in F$ then for some U containing p, $K(p) \cap U \subset F$. Then F is extremal.*

In fact, if $p \in F$ is not a critical point of $f = d_q$ then p is a regular point of f (cf. Remark 180). Then by Theorem 186, f is a trivial bundle map in a neighborhood of p. By moving along the slice of $\mathbb{R}$ through p we can then find points $a \in K(p)$ (and hence in F) arbitrarily close to p such that $f(a) < f(p)$; so p cannot be a local minimum of f on F. Since the strata of an MCS-space are determined by the topological type of conical neighborhoods of their points, we immediately have:

COROLLARY 191. *The strata of the stratification of X from Theorem* 168 *are extremal.*

Extremal sets have the following useful property. Note that there are variations in the literature on the following definition and notation:

DEFINITION 192. Let F be a closed subset of X and $p \in F$. The closure in Σ_p of all directions of minimal curves from p to F is denoted by $\Sigma_p F$.

PROPOSITION 193. *A closed set F in X is extremal if and only if $\Sigma_p F$ is extremal for all $p \in F$. (However, the "if" part may not be valid if we include the additional conditions for positive curvature in Definition* 187.)

One of the points of [66] is to show that Theorem 186 remains valid for the restrictions of admissible functions to extremal sets, and here a key point is that the observation made in Remark 184 is no longer valid. This problem is overcome by introducing the more general class of $\widetilde{MCS}$-spaces, where the conical neighborhoods are permitted to be isomorphic to cones over compact $\widetilde{MCS}$-spaces of dimension $\leqslant n-1$ instead of dimension equal to $n-1$. In consequence, the Theorems 168 and 167 also hold for extremal sets.

PROPOSITION 194. *For every compact set $K \subset X$ there exists an $\varepsilon > 0$ such that if $F \subset K$ is extremal and d_F denotes the induced inner metric on F (Section* 4.1), *then for every $x, y \in F$ $d_F(x, y) \leqslant \frac{1}{\varepsilon} d(x, y)$.*

A natural question to ask at this point is whether an extremal set, with the induced inner metric, is again an Alexandrov space with the same curvature bound. A counterexample to this statement for a codimension 3 extremal subset is given in [70]. The question remains open for smaller codimension. The next statements are also proved in [66] (the first part of the next proposition is obvious from the definition):

PROPOSITION 195. *If F and G are extremal subsets of X with $F \neq G$ then $F \cup G$, $F \cap G$ and $\overline{F - G}$ are all extremal sets.*

DEFINITION 196. An extremal subset F in X is called *primitive* if for any other extremal set G in X, $F \cap G$ is not open in F.

PROPOSITION 197. *If F is extremal in X then F can be written uniquely as a locally finite union of primitive subsets each having nonempty intersection with F.*

Extremal sets occur naturally in connection with quotients (cf. Section 4.4).

PROPOSITION 198. *Suppose Γ is a compact group acting on X by isometries, and $\pi : X \to G/\Gamma$ is the quotient map. Then*

(1) *if F is extremal in X then $\pi(F)$ is extremal in G/Γ, and*
(2) *if Γ' is a closed subgroup of Γ and F is the set of all fixed points of Γ' then $\pi(F)$ is extremal in G/Γ.*

We mention here a useful construction in the proof of this proposition, namely the convex hull of a compact set. We state a special case of this result, which is given a separate (and simpler) proof in [46].

PROPOSITION 199. *For every $x \in X$ there are a neighborhood U of x and a strictly concave function $f : U \to \mathbb{R}$ so that x is the maximum of f and the inverse images of points are compact.*

The construction of f is such that it is "stable" in a sense under Gromov–Hausdorff convergence.

9.8. *Quasigeodesics and gradient curves*

Note that X itself is obviously extremal; therefore the following statements apply to X as a whole. The existence of quasigeodesics in an Alexandrov space was proved in [65].

THEOREM 200. *Let $F \subset X$ be extremal, $p \in F$, and $\gamma \in \Sigma_p F$, and f be κ-convex without critical points on a neighborhood of F. Then*

(1) *any gradient curve which begins in F remains in F,*
(2) *there exists a quasigeodesic χ defined on $\mathbb{R}^+$ starting at p, with $\chi'(0) = \gamma$.*

In fact, $\frac{\nabla f}{|\nabla f|} \in \Sigma_p F$ whenever $p \in F$. We remind the reader of our comments after Proposition 147, which describe the construction of a "gradient exponential map".

The following has been called a "Generalized Lieberman Lemma", after the result in [52]. It was proved first in [66], and again with a simpler proof in [70].

THEOREM 201. *Let $F \subset X$ be extremal and let d_F denote the induced inner metric on F. Then minimal curves in d_F are quasigeodesics in X.*

9.9. *Boundary*

DEFINITION 202. If X is a 1-dimensional Alexandrov space, then the *boundary points* ∂X of X are the endpoints of X if X is an interval; otherwise X has no boundary. If X is an Alexandrov space of dimension $n \geqslant 2$ then $p \in X$ is a boundary point if and only if Σ_p contains boundary points. A point which is not a boundary point is called an *interior point*.

DEFINITION 203. We now establish some notation we will use from now on for Alexandrov spaces of curvature $\geqslant 1$, which will always denote by Σ. If Σ is 0-dimensional, then Σ is a point if $\partial\Sigma \neq \emptyset$, and a pair of points at distance π otherwise. If Σ is 1-dimensional then Σ is a closed interval of length π if $\partial\Sigma \neq \emptyset$, and a circle of diameter π otherwise.

Then next theorem was proved in [62], by induction on the dimensions of the two spaces.

THEOREM 204. *If Σ, Σ' are Alexandrov spaces of curvature $\geqslant 1$ (and possibly different dimension) such that $\mathbb{R}^k \times c\Sigma$ is homeomorphic to $\mathbb{R}^m \times c\Sigma'$ for some m, k, then Σ is with boundary if and only if Σ' is with boundary.*

COROLLARY 205. *A point $p \in X$ is a boundary point if and only if X has a conical neighborhood isomorphic to the cone on an Alexandrov space with boundary. In particular, the boundary is closed.*

From Proposition 190 we now get:

COROLLARY 206. *The boundary of an Alexandrov space is extremal.*

COROLLARY 207. *If $p \in \partial X$ then p has a conical neighborhood in ∂X isomorphic to $c\partial\Sigma_p$.*

A following "Gluing Theorem" was proved originally in the 2-dimensional case by Alexandrov. Perelman [62] proved a related "Doubling Theorem" for doubling an Alexandrov space with boundary. The generalization here is due to Petrunin [70].

THEOREM 208. *Let X_1 and X_2 be Alexandrov spaces of curvature $\geqslant k$ such that there is an isometry $\phi : \partial X_1 \to \partial X_2$ (where each boundary is given the induced inner metric). Then the space obtained by gluing X_1 and X_2 together along their boundaries via ϕ naturally has the structure of an Alexandrov space of curvature $\geqslant k$ having no boundary.*

In particular, one can "double" a space with boundary by using two copies of it and the identity map on the boundary.

10. Differentiable structures

Throughout this section X denotes an Alexandrov space of dimension n. This applies in particular to our comments about spaces with curvature above. Although it is possible to consider an upper curvature bound without a lower curvature bound (and this is even essential for, say, the theory of nonpositively curved groups), such an approach is outside the scope of this article. A good reference for the upper curvature bound condition is [4].

10.1. *Curvature bounded below and above*

As mentioned in the introduction, Alexandrov's problem of synthetic differential geometry was solved by Berestovskii and Nikolaev [6,55–57]. Their solution involved spaces with curvature bounded both above and below. Curvature bounded above in an Alexandrov space can be defined by reversing the inequalities in any of the three comparison conditions A1, A2, and A3 (cf. Section 3.3), and assuming that one (hence all) of these conditions is true locally. Despite the analogous conditions, there are significant differences between curvature bounded above and curvature bounded below. For example, the Global Comparison Theorem (Theorem 43) is *not* valid for curvature bounded above. For example, the flat torus $S^1 \times S^1$ has curvature locally bounded above and below by 0 (so curvature $= 0$) locally, but not globally. The easiest way to see this is to note the following fundamental property of curvature $\leqslant K$. Since the space forms S_K have the property that geodesics starting at a point p of length less than $\pi/\sqrt{K}$ do not meet, the same must be true in a region of curvature $\leqslant K$ in X. Thus no compact space can have globally curvature $\leqslant 0$. It also follows that for any point p, the cut radius map c_p (Definition 107) has a positive lower bound. If in addition space is geodesically complete (Definition 25) then following result is immediate from our previous work on Alexandrov spaces:

THEOREM 209. *Suppose X has curvature $\leqslant K$ and is geodesically complete. Then for every $p \in X$*
(1) *p is regular,*
(2) *$T_p = \mathbb{R}^n$,*
(3) *$\exp_p$ is a local homeomorphism.*

From the third part of the above statement it follows that X is a topological manifold. Berestovskii introduced the distance coordinates to prove that X is a differentiable manifold with continuous Riemannian metric. Later, Nikolaev synthetically constructed a notion of parallel transport, which he used to build so-called harmonic coordinates. He proved the following theorem. Note that finite dimensionality is not a required assumption.

THEOREM 210. *If Y is a geodesically complete inner metric space of curvature bounded above and below then Y has the structure of a smooth manifold, and the distance d on Y is induced by a Riemannian metric on Y, which has class $c^{1,\alpha}$ with respect to the harmonic coordinates.*

Nikolaev later showed [13] that the smoothness of the Riemannian metric cannot be improved; that is, the above theorem is the best possible solution to Alexandrov's problem of synthetic differential geometry that can be obtained using curvature bounds. Nikolaev also proved [58] that every geodesically complete space of curvature bounded above and below can be approximated Riemannian manifolds having nearly the same (sectional) curvature bounds and the same dimension. The approximation is in the Gromov–Hausdorff sense, and in the stronger sense of bi-Lipschitz homeomorphisms.

The following theorem generalizes Berestovskii's result to spaces which are not geodesically complete [73]. Note that many issues already arise even when only the geodesic completeness condition is dropped, in particular the problem of finite dimensionality. For example, the Hilbert cube $[0, 1] \times [0, 1/2] \times \cdots$ clearly is compact and has curvature $= 0$, but is infinite dimensional.

THEOREM 211. *Suppose Y is an inner metric space of curvature bounded above and below. Then the following are equivalent:*

(1) *the set $\mathcal{T}$ of geodesic terminals (Definition 108) in Y is nowhere dense,*

(2) $\dim Y < \infty$,

(3) *Y is homeomorphic to a smooth manifold with boundary, and $\partial Y = \mathcal{T}$.*

Since a smooth manifold with boundary admits a Riemannian metric such that the induced metric has curvature bounded below, including the boundary, one has:

COROLLARY 212. *A topological space M admits the structure of a smooth manifold with boundary if and only if it is finite dimensional and has a metric of curvature bounded above and below.*

10.2. *Other differentiable structures*

As mentioned in the previous section, curvature bounded above implies that the cut radius at every point has a positive lower bound. One can therefore weaken the notion of an upper curvature bound by requiring that the space have *positive cut radius* (in Riemannian geometry this means the same thing as having positive "injectivity radius"). Note that a differentiable manifold may have several nondiffeomorphic differentiable structures. The next theorem was proved in [74]:

THEOREM 213. *If X positive cut radius then X is a smooth manifold having a unique differentiable structure such that d_p is differentiable near (but not at) p.*

In more general cases there are differentiable structures, but not uniqueness results. The next theorem was proved in [60] and [59]. We refer the reader to those papers for the precise definitions. The general theory of differentiable structures on Alexandrov spaces seems still to be in a state of development.

THEOREM 214. *Let X be an Alexandrov space.*

(1) *There exists a set X_0 such that $X - X_0$ has (n-dimensional) measure 0 and contains the set S_X of singular points in X.*

(2) *The metric on X comes from a $c^{1/2}$-Riemannian metric g on X_0.*

(3) *g extends continuously to $X - S_X$.*

(4) *X has an almost everywhere approximately second differentiable structure in the sense of Stolz.*

11. Convergence and stability

Throughout this section we will use the conventions described in Section 4.2. X will always denote an Alexandrov space.

11.1. *Convergence*

Let $\mathcal{M}(k, n, D)$ denote the class of all Alexandrov spaces of curvature $\geqslant k$, dimension $\leqslant n$, and diameter $\leqslant D$, endowed with the Gromov–Hausdorff metric. In this section we study this class, showing that it is compact, and considering the relationship between the elements of a sequence of spaces in this class and its limit.

THEOREM 215. *The space $\mathcal{M}(k, n, D)$ is compact in the Gromov–Hausdorff metric.*

PROOF. We know already from Lemma 152 and Theorem 43 that $\mathcal{M}(k, n, D)$ is closed. For precompactness we follow the proof of Theorem H in [72]. We can assume $k < 0$, and, by scaling, we can assume $k = -1$. Fix $\varepsilon > 0$ and let $\mathcal{N}'$ be an $\varepsilon/2$-net in $\overline{B}(0, D) \subset S^n_{-1}$, having m elements. Let $X \in \mathcal{M}(k, n, D)$ and $p \in X$ be a regular point. Let T_p have the elliptic cone metric (Section 4.5) so that T_p is isometric to S^n_{-1} and $\exp_p$ is distance nonincreasing by Theorem 43. Consider $C = \exp_p^{-1}(X) \cap \overline{B}(0, D)$, and choose a set $\mathcal{N}$ as follows. For each $x' \in \mathcal{N}'$, if $B(x', \varepsilon/2) \cap C \neq \emptyset$, choose a point x in $B(x', \varepsilon/2) \cap C$. The resulting set is clearly an ε-net in C having fewer than m elements. Since $\exp_p$ is distance nonincreasing, the image of this set in X is an ε-net having fewer than m elements. □

As pointed out in Section 4.2, in studying a Gromov–Hausdorff convergent sequence one can always assume that the sequence has been embedded in some ambient metric space, and the convergence is in the usual Hausdorff sense. In particular, we can make sense of uniform convergence of curves, and statements proved earlier about convergence of minimal curves, semicontinuity of angles, etc., are all valid in terms of Gromov–Hausdorff convergence. For example, Lemma 36 and Proposition 57 can be restated as follows:

PROPOSITION 216. *Suppose $X_i \in \mathcal{M}(k, n, D)$, with $X_i \to X$, and we have $p_i \in X_i$ with $p_i \to p \in X$. If $\gamma_i, \beta_i \in S_{p_i}$ and $\gamma_i \to \gamma$ and $\beta_i \to \beta$, where $\gamma, \beta \in S_p$, then $\alpha(\gamma, \beta) \leqslant \liminf \alpha(\gamma_i, \beta_i)$. If γ is almost extendable past p then $\alpha(\gamma, \beta) = \liminf \alpha(\gamma_i, \beta_i)$.*

11.2. *Stability theorems*

A very important question is, if $X_i \to X$, what is the topological relationship between X_i and X for large i? In complete generality there is no relationship. Any compact inner metric space can be scaled by constants converging to 0 to be made arbitrarily close to a point. In $\mathcal{M}(k, n, D)$ a strong answer is given in [62] when there is no "collapse" in the limit:

THEOREM 217. *If* $X \in \mathcal{M}(k, n, D)$ *then there is a neighborhood* $\mathcal{U}$ *of* X *such that if* $Y \in \mathcal{U}$ *and* $\dim Y = n$, *then* Y *is homeomorphic to* X.

The proof of the theorem is quite technical, and is based on the Deformation Theorem 5.4 of [83]. Since $\mathcal{M}(k, n, D)$ is compact, we immediately get the following:

COROLLARY 218. *The class* $\mathcal{M}^*(k, n, D)$ *of all* n*-dimensional elements of* $\mathcal{M}(k, n, D)$ *has finitely many homeomorphism types. More precisely, there exist* $X_1, \ldots, X_k \in \mathcal{M}^*(k, n, D)$ *such that every* $Y \in \mathcal{M}^*(k, n, D)$ *is homeomorphic to some* X_i.

The above corollary is an example of a so-called "finiteness theorem". The first completely general such theorem in Riemannian geometry is Cheeger's Finiteness Theorem [21].

COROLLARY 219. *For any* k *and positive numbers* D, v, *the class of Riemannian* n*-manifolds having sectional curvature* $\geqslant k$, *diameter* $\leqslant D$, *and volume* $\geqslant v$ *contains finitely many homeomorphism types for any* n, *and finitely many diffeomorphism types for* $n \neq 4$.

This is the well-known finiteness theorem of Grove–Petersen–Wu [43] (except they did not prove homeomorphism finiteness for dimension 3). It follows easily from Theorem 217 because volume in the Riemannian sense and n-dimensional Hausdorff measure differ by a multiplicative positive constant. So a uniform lower bound on volume prevents Hausdorff measure from going to zero, and hence Hausdorff dimension from dropping. Note also that the diffeomorphism part of the theorem follows from well-known work of Kirby and Seibenmann. Note that in [43] a more general theorem is proved for so-called $LGC(\rho)$ manifolds, which are more general than Alexandrov spaces. On the other hand, Theorem 217 is valid for spaces which are not manifolds.

Another precursor of Theorem 217 is the so-called Gromov Compactness Theorem [31,67]. As we pointed out in Proposition 158, the tangent cone is the pointed Gromov–Hausdorff limit of the scalings of any small neighborhood of p. By using a more general statement of Theorem 217 one gets:

COROLLARY 220. *If* $x \in X$ *then* x *has a spherical neighborhood pointed homeomorphic to the tangent cone* T_p.

COROLLARY 221. *If* X *has curvature* $\geqslant 1$ *and* $\operatorname{diam}(X) > \pi/2$ *then* X *is homeomorphic to the suspension of a compact* $(n-1)$*-dimensional Alexandrov space of curvature* $\geqslant 1$.

Since a Riemannian manifold (being a topological manifold) is locally homeomorphic to a cone on a lower dimensional sphere, in case X is a Riemannian manifold the conclusion can be strengthened to X is homeomorphic to a sphere. Therefore the corollary is a generalization of the Grove–Shiohama Diameter Sphere Theorem. In fact, we can reach the same conclusion by supposing that X is geodesically complete – for then the space of directions at every point is a sphere. To prove the corollary, let p and q be such that $d(p, q) = \operatorname{diam}(X) > \pi/2$. It follows from the geometry of the sphere that for every

$x \neq p, q$, $\alpha_1(x; p, q) > \pi/2 + \varepsilon$ for some $\varepsilon > 0$. It follows from Proposition 123 that x is never a critical point of d_p. By Theorem 186(3), X is homeomorphic to the suspension of the boundary ∂B of any metric ball B centered at p. By Corollary 220 and the uniqueness of conical neighborhoods, we get that ∂B is homeomorphic to Σ_p, which is a space of curvature $\geqslant 1$.

In case the spaces X_i have a fixed positive lower bound on cut radius, it is possible to prove stability up to diffeomorphism directly (cf. [74], and a correction in [90]). Let $\mathcal{M}(k, n, D, \varepsilon)$ denote the subset of $\mathcal{M}(k, n, D)$ consisting of all elements having cut radius $\geqslant \varepsilon > 0$ at every point, endowed with their unique compatible differentiable structures (Theorem 213). This theorem generalizes Cheeger's Finiteness Theorem:

THEOREM 222. *If $X \in \mathcal{M}(k, n, D, \varepsilon)$ then there is a neighborhood $\mathcal{U}$ of X such that if $Y \in \mathcal{U}$ and* $\dim Y = \dim X$, *then Y is diffeomorphic to X.*

Obviously, if $X_i \to X$, and $\dim X < \dim X_i$, X cannot be homeomorphic to X_i. In this situation, which is referred to as *collapsing* , the best that one can hope for is some kind of bundle map $X_i \to X$ with nice fibers. This situation has been studied extensively in Riemannian geometry (cf. [28,91,24]). There are two results along these lines for Alexandrov spaces, [90] and [92]. The first theorem requires positive cut radius and the second also requires a kind of regularity.

11.3. *Approximation*

It is natural to ask whether spaces in $\mathcal{M}(k, n, D)$ can be approximated in the Gromov–Hausdorff sense by Riemannian manifolds having sectional curvatures uniformly bounded below, as is true with an upper curvature bound and geodesic completeness [58]. Obviously, by Theorem 217, a space which is not a topological manifold cannot be approximated by Riemannian manifolds of the same dimension. It is currently an open question whether elements of $\mathcal{M}(k, n, D)$ can be approximated by Riemannian manifolds of higher dimension and curvature uniformly bounded below. In [68], however, it is shown that it is not always possible to approximate Alexandrov spaces of curvature $\geqslant 1$ by Riemannian manifolds of sectional curvature $\geqslant 1$ of any dimension. The examples in question are suspensions of Riemannian manifolds of curvature $\geqslant 1$.

12. Metric invariants and recognition

12.1. *Recognition problem*

The "Recognition Problem" (cf. [36]) for a given class $\mathcal{M}$ of metric spaces, such as Alexandrov spaces or Riemannian manifolds, asks to what degree can a space be identified, up to homotopy-type, homeomorphism, diffeomorphism or isometry, using finitely many metric measurements – i.e., *metric invariants*, which assign to any metric space a real number that depends only on its isometry type. So far we have used the invariants diameter, cut radius, Hausdorff measure ("volume") and dimension, and curvature bounded below and above. Given a collection $I_1, \ldots, I_k$ of such invariants one obtains a map $\mathcal{I}: \mathcal{M} \to \mathbb{R}^k$. There are three natural kinds of problems to consider:

(1) The *range problem*: identify $\mathcal{I}(\mathcal{M})$.
(2) *Slice problems*: recognize spaces in $\mathcal{I}^{-1}(\Omega)$, $\Omega \subset \mathbb{R}^k$.
(3) *Extremal problems*: recognize spaces X with $\mathcal{I}(X)$ on or close to the boundary of $\mathcal{I}(\mathcal{M})$.

Corollary 47 is an example of a solution to the range problem. If we let x be the lower curvature bound and y be diameter, and consider the class of complete inner metric spaces, then $\mathcal{I}(\mathcal{M})$ is all of the second quadrant, and all points below the graph of $\pi/\sqrt{x}$ in the first quadrant. If we further restrict $\mathcal{M}$ by requiring dimension $\leqslant n$ then Corollary 221 is a solution to a slice problem. The following proposition, which is easily proved using Proposition 35, is an example of a solution to the extremal problem (note that by scaling it is valid for any positive curvature bound):

PROPOSITION 223. *If X is an inner metric space of curvature $\geqslant 1$ and diameter π, then X is isometric to the the suspension $\Sigma_1 Y$ of a space of curvature $\geqslant 1$. In particular, if X is geodesically complete then X is isometric to a standard sphere.*

In Riemannian geometry, a major category of solutions to extremal and slicing problems are the so-called sphere theorems. It is a long-standing open question whether there are exotic spheres (i.e., differentiable manifolds homeomorphic, but not diffeomorphic, to the standard sphere) having positive sectional curvature (equivalently in the sense of Alexandrov). A related question is also open: can the Grove–Shiohama Sphere Theorem be improved to diffeomorphism type; that is, do there exist exotic spheres of curvature $\geqslant 1$ and diameter $> \pi/2$? Alexandrov spaces have played a key role in recent progress on this question, which we will describe below.

12.2. *Some metric invariants*

DEFINITION 224. The *radius* of a metric space Y is the quantity

$$\operatorname{rad}(Y) := \min_{x \in Y}\Big\{\max_{y \in Y}\{d(x, y)\}\Big\}.$$

In other words, the radius is the smallest $r > 0$ such that Y is entirely contained in $\overline{B}(x, r)$ for some $x \in Y$. If $\operatorname{rad}(Y) > r$ then for every $x \in Y$ there exists a $y \in Y$ such that $d(x, y) > r$. Also, $\operatorname{rad}(Y) \leqslant \operatorname{diam}(Y) \leqslant 2\operatorname{rad}(Y)$.

DEFINITION 225. For $q \geqslant 2$, the qth *packing radius* of a compact metric space Y is the quantity

$$\operatorname{pack}_q(Y) := \frac{1}{2} \max_{x_1, \ldots, x_q \in Y}\Big\{\max_{1 \leqslant i \leqslant j \leqslant q} d(x_i, x_j)\Big\}.$$

If $\{x_1, \ldots, x_q\}$ realizes $\operatorname{pack}_q(Y)$ in the above formula then $\{x_1, \ldots, x_q\}$ is called a *q-packer.*

In other words, $\operatorname{pack}_q(Y)$ is the largest $r > 0$ such that Y contains q disjoint open r-balls. Then we have

$$\frac{1}{2}\operatorname{diam}(Y) = \operatorname{pack}_2(Y) \geqslant \operatorname{pack}_3(Y) \cdots$$

and $\lim_{q\to\infty} \operatorname{pack}_q(Y) = 0$.

The last invariant is from [36], where also the notion of *q-extent* $xt_q(Y)$ is defined as the maximal average distance between q-tuples of points in Y. Two related invariants are defined in [86]: *q-covering radius* $\operatorname{cov}_q(Y)$, the smallest $\varepsilon > 0$ such that there is an ε-net of q elements in Y; and *q-paving radius* $\operatorname{pav}_q(Y)$, the smallest $r > 0$ such that Y is a union of q subsets having diameter at most $2r$. We are unable to discuss the results of [36] and [86] in great detail here; however these papers are very readable.

12.3. *Sphere theorems*

"Sphere theorems" are theorems which allow one to conclude from measuring certain metric invariants that the space in question is a sphere – up to homotopy-type, homeomorphism, diffeomorphism, isometry (or some variation such as bi-Lipschitz homeomorphism). Sphere theorems naturally fall into two classes: "hard" and "soft". Hard sphere theorems are ones in which the geometric conditions force the space in question to be almost isometric to a standard sphere. Often they are proved applying a stability theorem like Theorem 217. For example, it follows from Theorem 149 that if X has (dimension n and) curvature $\geqslant 1$ and contains a spherical set of $2(n+1)$ elements then X is isometric to S_1^n. If one takes a sequence of spaces X_i of dimension n and curvature $\geqslant 1$ and sets of $2(n+1)$ elements that are closer and closer to being spherical in any reasonable sense, then the sequence must Gromov–Hausdorff approach the standard sphere. One can then apply known convergence results for Riemannian manifolds, such as [91], to obtain that a Riemannian manifold with curvature $\geqslant 1$, containing an "almost" spherical set, is diffeomorphic to a standard sphere. See [76, Theorems 4.7 and 4.9].

Soft sphere theorems (which generally seem harder to prove!) involve geometric restrictions that do not force the space to be close to a sphere, and therefore cannot be proved directly with a stability theorem. We list some examples of both kinds here. We have already mentioned the Grove–Shiohama Sphere Theorem, which is soft, because any suspension of a space of curvature $\geqslant 1$ has diameter π, but may not be a sphere. The next theorem was proved independently in [42] and [70].

THEOREM 226. *If X is an inner metric space of curvature $\geqslant 1$ and* $\operatorname{rad}(X) > \pi/2$ *then X is homeomorphic to S^n.*

The proof in [70] is somewhat easier and shows the utility of quasigeodesics. One first shows that at every point $p \in X$, Σ_p has radius $> \pi/2$. If not, then there exists a direction

$\gamma \in \Sigma_p$ such that $\overline{B}(\gamma, \pi/2) = \Sigma_p$. Take a quasigeodesic χ of length $\pi/2$ tangent to γ, which ends at a point q. If x is any other point in X and γ is minimal from x to p then $\alpha(\gamma, \chi) \leqslant \pi/2$, and it follows (since quasigeodesics satisfy the usual comparisons) that $d(q, x) \leqslant \pi/2$. But this contradicts $\mathrm{rad}(X) > \pi/2$. The proof of the theorem now proceeds by induction on dimension. Suppose it is true for $\dim X \leqslant n-1$. Then since $\mathrm{diam}(X) \geqslant \mathrm{rad}(X) > c\pi/2$ then Corollary 221 implies that X is the suspension of an Alexandrov space Σ of curvature $\geqslant 1$. By the uniqueness of conical neighborhoods we must have $\Sigma = \Sigma_p$, which, by what we have proved and the inductive hypothesis, is a sphere.

We now mention a soft diffeomorphism sphere theorem for Riemannian manifolds from [45,46]. Although it is not a sphere theorem for Alexandrov spaces (which need to be homeomorphic to spheres in this class), we mention it because Alexandrov spaces come into the proof in a novel way, as we will explain below.

THEOREM 227. *If $n \geqslant 4$ then any closed Riemannian n-manifold M of sectional curvature $\geqslant 1$ and* $\mathrm{pack}_{n-2}(M) > \pi/4$ *is diffeomorphic to S^n.*

Note that we already know from the Grove–Shiohama Theorem that M is homeomorphic to a sphere, and that M is diffeomorphic to a sphere for $n \leqslant 6$. For $n \geqslant 5$ it follows from the h-cobordism theorem, Lemma 2.3 of [49] and Theorem 9.7 in [53] (Brown's Theorem) that we need only exhibit a smooth embedding of S^n into $\mathbb{R}^{n+1}$. A key point in the proof is the reduction of the global problem of understanding the diffeomorphism type of M to a local problem in an Alexandrov space. This is done by considering the suspension $\Sigma_1 M$, which we know is an Alexandrov space. The suspension of a Riemannian manifold can also be described as the warped product [17] $M \times_f (0, \pi)$, where $f(t) = \sin t$, with two "endpoints" attached. Therefore $\Sigma_1 M$ minus the endpoints is in fact a Riemannian manifold diffeomorphic to $M \times \mathbb{R}$. The idea, then, is to use Alexandrov space techniques to construct a mapping of a small anulus near an endpoint $*$ of $\Sigma_1 M$ into a Euclidean space. In case $\mathrm{pack}_{n+2} M > \pi/4$ the problem is relatively simple. Choose $n+2$ pairs of points in M at pairwise distance $> \pi/2$; these correspond to directions γ_i at $*$. The distance functions from points on the first $n+1$ geodesics γ_i then together form an admissible map into $\mathbb{R}^{n+1}$, with $*$ (and hence nearby points) as regular points. We can then apply Theorem 186 to get a topological embedding into $\mathbb{R}^{n+1}$. The mapping can be made smooth using standard averaging techniques. The improvement to pack_{n-2} requires more careful arguments.

12.4. *Related results*

A metric characterization of locally symmetric spaces, and a "pinching" theorem for such spaces is given in [75]. Isometry groups of Alexandrov spaces are considered in [30]. The solution for a slice problem for the invariants lower curvature bound, radius, and volume for Riemannian manifolds may be found in [41]. The volume comparison of [41] is generalized to Alexandrov spaces in [85]. Other examples of "hard" pinching theorems for Riemannian manifolds may be found in, e.g., [61]. Further geometric constraints on exotic spheres of positive curvature may be found in [46]. In [81], it is proved that a Finsler manifold with a lower curvature bound is Riemannian.

13. Infinite dimensional spaces

Beyond Theorems 27, 43 and 151, very little is known in general about infinite dimensional spaces of curvature bounded below, and progress seems difficult. It is not known whether the space of directions is an inner metric space at every point; the best regularity known is contained in the following immediate consequence of Theorem 151 (see Definition 150 for notation):

THEOREM 228. *If X is an inner metric space of curvature bounded below, then for every n and $p \in X$, $R^n(X)$ is dense in X.*

To make further progress seems to require assumptions about geometric regularity, such as homogeneity (i.e., X has a transitive group of isometries). A finite dimensional homogeneous space of curvature bounded below is isometric to a Riemannian homogeneous space. This fact, which is stated without proof in [19], is reasonable, given that the tangent space at every point must be Euclidean. However, we do not know of an easy rigorous proof of this fact. One can proceed as follows: according to Berestovskii, a finite dimensional homogeneous inner metric space is isometric to a homogeneous manifold with invariant Carnot–Caratheodory metric [11]. The canonical example of a homogeneous space is a topological group with left-invariant metric; the group acts transitively and effectively on itself by isometries via left multiplication. We summarize the main known results about infinite dimensional homogeneous spaces of curvature bounded below, taken from [77,15, 16]. We denote the identity of a group by e; a 1*-parameter subgroup* is a homomorphism from the reals into G.

THEOREM 229. *Every locally compact, metrizible, arcwise connected topological group G admits a left invariant, geodesically complete inner metric of curvature bounded below. Furthermore,*

(1) *If G is compact, the metric can be assumed bi-invariant.*
(2) *Every direction in Σ_p is tangent to a unique* 1*-parameter subgroup.*

Note that metrizibility and arcwise connectedness are necessary for the existence of any inner metric on G. The theorem can be proved either by approximation by Lie groups (cf. [77]), or in the following way (cf. [16]): G has a "universal cover" that splits as a product $\widetilde{G} = L \times G_1 \times \cdots \times \mathcal{R}^{\omega}$, where L is a Lie group, each G_i is a compact simple Lie group, and $\mathcal{R}^{\omega}$ is a product of reals, where ω may be finite or countably infinite (any of these factors may be trivial; if L is nontrivial it is not compact.) The "covering map" $\pi : \widetilde{G} \to G$ factors as a composition $\pi' : \widetilde{G} \to L \times K$, where K is a compact group, followed by $\pi'' : L \times K \to G$, where π'' is a covering map in the usual sense. There is a surjective homomorphism from $\widetilde{K} := G_1 \times \cdots \times S^1 \times \cdots \to K$, where $S^1 \times \cdots = T^{\omega}$ is a (possibly infinite) torus. These compact factors of $\widetilde{K}$ are given bi-invariant Riemannian metrics scaled so that their diameters are square summable. (The curvature of bi-invariant Riemannian metrics is nonnegative, so scaling preserves that lower curvature bound). $\widetilde{K}$ is given the product metric, and so has a bi-invariant metric of curvature bounded below, and K is given the quotient metric. If G was compact, $G = K$, and we are finished. Otherwise, we give L a left-invariant Riemannian metric, $L \times K$ the product metric, and G the quotient metric. We have to pass to $\widetilde{K}$ because $\mathcal{R}^{\omega}$ does not admit an invariant inner metric [16].

We have the following corresponding theorem for homogeneous spaces [15]:

THEOREM 230. *Every locally connected quotient G/H of a locally compact, connected, first countable group G by a compact subgroup H admits a metric of curvature $\geqslant k$, for some k, invariant under the action of G. If G/H is compact, we can take $k = 0$.*

There are a number of theorems generalizing classical theorems for Riemannian homogeneous spaces. An inner metric is called *weakly flat* if for any two geodesics γ, β starting at a point p, $d(\gamma(t), \beta(t)) = 2t\sin(\alpha(\gamma, \beta)/2)$ for all sufficiently small $t > 0$; in other words γ, β span a small Euclidean triangle. The torus in Example 101 is weakly flat. A weakly flat Riemannian manifold is flat.

THEOREM 231. *Let G be a group with bi-invariant metric of curvature bounded below. Then*

(1) *G has nonnegative curvature.*
(2) *Every geodesic through e is a 1-parameter subgroup.*
(3) *G is abelian if and only if the metric of G is weakly flat.*
(4) *If G is locally compact then G is a direct product (algebraic and metric) of a compact group with bi-invariant metric of curvature $\geqslant 0$ and a finite dimensional Euclidean vector space and*
(5) *If G is compact and abelian then G is isometric to a flat Riemannian torus or a weakly flat infinite torus.*

There are some corresponding results for so-called *normal* homogeneous spaces. In addition we have:

THEOREM 232. *Every locally compact nonnegatively curved homogeneous inner metric space is isometric to the direct metric product of a finite dimensional Euclidean space and a compact nonnegatively curved homogeneous space.*

THEOREM 233. *Every locally compact homogeneous inner metric space with curvature $\geqslant k > 0$ must be a (finite dimensional) Riemannian homogeneous space with sectional curvature $\geqslant k$.*

Infinite dimensional groups with invariant metrics of curvature bounded below have some properties unlike those of Lie groups with invariant Riemannian metrics. For example, there exist nonrectifiable 1-parameter subgroups. In fact, even in a weakly flat torus (see Example 101) there are 1-parameter subgroups having no rectifiable curves in their homotopy class (cf. [16]), contradicting a statement of Gromov [32] that in a compact inner metric space, every curve has a geodesic in its homotopy class. Generally, one cannot find a neighborhood of e in which every point is uniquely joined to e by either a minimal curve or a 1-parameter subgroup.

References

[1] A.D. Alexandrov, *Quasigeodesics*, RAN USSR **69** (6) (1949), 717–720 (Russian).
[2] A.D. Alexandrov, *Die innere Geometrie der konvexen Flächen*, Akad. Verl., Berlin (1955).
[3] A.D. Alexandrov, *Über eine Verallgemeinerung der Riemannschen Geometrie*, Schr. Forschungsinst. Math. **1** (1957), 33–48.
[4] A.D. Alexandrov, V.N. Berestovskii and I.G. Nikolaev, *Generalized Riemannian spaces*, Russian Math Surveys **41** (1986), 1–54.
[5] A. D. Aleksandrov and V.A. Zalgaller, *Intrinsic Geometry of Surfaces*, Transl. Math. Monogr., Vol. 15, Amer. Math. Soc., Providence, RI (1967).
[6] V.N. Berestovskii, *Introduction of a Riemannian structure in certain metric spaces*, Siberian Math. J. **16** (1977), 210–221.
[7] V.N. Berestovskii, *Spaces of bounded curvature*, Dokl. Akad. Nauk SSSR **258** (1981), 269–271.
[8] V.N. Berestovskii, *Homogeneous Busemann G-spaces*, Siberian Math. J. **23** (1982), 3–15.
[9] V.N. Berestovskii, *Borsuk's problem on metrization of a polyhedron*, Dokl. Akad. Nauk SSSR **268** (2) (1983), 273–277 (Russian).
[10] V.N. Berestovskii, *Spaces with bounded curvature and distance geometry*, Siberian Math. J. **27** (1986), 8–19.
[11] V.N. Berestovskii, *Homogeneous spaces with intrinsic metric*, Soviet Math. Dokl. **27** (1989), 60–63.
[12] V.N. Berestovskii, *Homogeneous manifolds with intrinsic metric II*, Siberian Math. J. **30** (1989), 180–190.
[13] V.N. Berestovskii, *On the structure of locally compact spaces with intrinsic metrics*, Siberian Math. J. **30** (1989).
[14] V.N. Berestovskii, *Homogeneous spaces with intrinsic metrics*, Dr. Sci. dissertation, Inst. Mat. Sib. Otdel. Akad. Nauk SSSR, Novosibirsk (1990).
[15] V.N. Berestovskii and C. Plaut, *Metric geometry of locally compact groups and quotients*, J. Geom. Anal. **9** (1999), 203–219.
[16] V.N. Berestovskii, C. Plaut and C. Stallman, *Geometric groups I*, Trans. Amer. Math. Soc. **351** (1999), 1403–1422.
[17] R.L. Bishop and B. O'Neill, *Manifolds of negative curvature*, Trans. Amer. Math. Soc. **145** (1969), 1–49.
[18] Yu. Burago, M. Gromov and G. Perel'man, *A.D. Alexandrov's spaces with curvatures bounded from below*, Preliminary version.
[19] Yu. Burago, M. Gromov and G. Perel'man, *A.D. Alexandrov's spaces with curvatures bounded from below*, Uspekhi Mat. Nauk **47** (2) (1992), 3–51 (in Russian), or Russian Math. Surveys **42** (1992), 1–58.
[20] H. Busemann, *The Geometry of Geodesics*, Academic Press, New York (1955).
[21] J. Cheeger, *Finiteness theorems for Riemannian manifolds*, Amer. J. Math. **92** (1970), 61–74.
[22] J. Cheeger and T. Colding, *On the structure of spaces with Ricci curvature bounded below; I*, Preprint.
[23] J. Cheeger and D. Ebin, *Comparison Theorems in Riemannian Geometry*, North-Holland Math. Library, Vol. 9, Amsterdam (1975).
[24] J. Cheeger and M. Gromov, *Collapsing Riemannian manifolds while keeping their curvatures bounded I*, J. Differential Geom. **23** (1986), 309–346.
[25] C. Chevalley, *Theory of Lie Groups*, Princeton University Press, Princeton, NJ (1946).
[26] S. Cohn-Vossen, *Existenz kurzester Wege*, Dokl. Math. SSSR **8** (1935), 339–342.
[27] M.P. do Carmo, *Riemannian Geometry*, Mathematics: Theory & Applications, Birkhäuser, Boston (1992).
[28] K. Fukaya, *A boundary of the set of Riemannian manifolds with bounded curvatures and diameters*, J. Differential Geom. **28** (1988).
[29] K. Fukaya and T. Yamaguchi, *The fundamental groups of almost nonnegatively curved manifolds*, Ann. of Math. **136** (1992), 253–333.
[30] K. Fukaya and T. Yamaguchi, *Isometry groups of singular spaces*, Math. Z. **216** (1994), 31–44.
[31] R. Greene and H. Wu, *Lipschitz convergence of Riemannian manifolds*, Pacific J. Math. **131** (1988), 119–141.
[32] M. Gromov, J. Lafontaine and P. Pansu, *Strutures métriques pour les variétés riemanniennes*, Cedic/Fernand Nathan, Paris (1981).
[33] K. Grove, *Metric differential geometry*, Differential Geometry, Lecture Notes in Math., Vol. 1263, Springer, Berlin (1987), 171–227.

[34] K. Grove, *Metric and topological measurements of manifolds*, Proc. Int. Cong. of Math. Kyoto, Japan, 1990, Springer, Berlin (1991), 511–519.
[35] K. Grove, *Critical point theory for distance functions*, Proc. Symp. Pure Math., Vol. 54 (1993), 357–385.
[36] K. Grove and S. Markvorsen, *New extremal problems for the Riemannian recognition program via Alexandrov geometry*, J. Amer. Math. Soc. **8** (1995), 1–28.
[37] K. Grove and P. Petersen, *Bounding homotopy types by geometry*, Ann. of Math. **128** (1988), 195–206.
[38] K. Grove and P. Petersen, *A pinching theorem for homotopy spheres*, J. Amer. Math. Soc. **3** (1990), 671–677.
[39] K. Grove and P. Petersen, *On the excess of metric spaces and manifolds*, Preprint.
[40] K. Grove and P. Petersen, *Manifolds near the boundary of existence* J. Differential Geom. **33** (1991), 379–394.
[41] K. Grove and P. Petersen, *Volume comparison á la Alexandrov*, Acta. Math. **169** (1992), 131–151.
[42] K. Grove and P. Petersen, *A radius sphere theorem*, Invent. Math. **112** (1993), 577–583.
[43] K. Grove, P. Petersen and J.-Y. Wu, *Geometric finiteness via controlled topology*, Invent. Math. **99** (1990), 205–214. *Erratum*, Invent. Math. **104** (1991), 221–222.
[44] K. Grove and K. Shiohama, *A generalized sphere theorem*, Ann. of Math. **106** (1977), 201–211.
[45] K. Grove and F. Wilhelm, *Hard and soft packing radius theorems*, Ann. of Math. **142** (1995), 213–237.
[46] K. Grove and F. Wilhelm, *Geometric constraints on exotic spheres via Alexandrov geometry*, J. Reine Angew. Math. **487** (1997), 201–217.
[47] W. Hurewicz and H. Wallman, *Dimension Theory*, Princeton University Press, Princeton (1941).
[48] K.W. Kwun, *Uniqueness of the open cone neighborhood*, Proc. Amer. Math. Soc. **15** (1964), 476–479.
[49] M. Kervaire and J. Milnor, *Groups of homotopy spheres: I*, Ann. of Math. **77** (1963), 504–537.
[50] C.-W. Kim and J.-W. Yim, *Surfaces with extreme value of curvature in Alexandrov spaces*, Tôhoku Math. J. **47** (1995), 461–474.
[51] U. Lang and V. Schroeder, *Kirszbraun's theorem and metric spaces of bounded curvature*, Geom. Funct. Anal. **7** (1997), 535–560.
[52] I.M. Lieberman, *Geodesics on convex surfaces*, RAN USSR **32** (5) (1941), 310–313 (Russian).
[53] J. Milnor, *Lectures on the H-Cobordism Theorem*, Princeton University Press (1965).
[54] D. Montgomery and L. Zippin, *Topological Transformation Groups*, Interscience, New York (1955).
[55] I.G. Nikolaev, *Parallel translation and smoothness of the metric of spaces of bounded curvature*, Soviet Math. Dokl. **21** (1980), 263–265.
[56] I.G. Nikolaev, *Parallel displacement of vectors in spaces of Alexandrov two-side-bounded curvature*, Siber. Math. J. **24** (1983), 106–119.
[57] I.G. Nikolaev, *Smoothness of the metric of spaces with bilaterally bounded curvature in the sense of A.D. Alexandrov*, Siberian Math. J. **24** (1983), 247–263.
[58] I.G. Nikolaev, *Bounded curvature closure of the set of compact Riemannian manifolds*, Bull. Amer. Math. Soc. **24** (1992), 171–177.
[59] Y. Otsu, *Almost everywhere existence of second differentiable structure of Alexandrov spaces*, Preprint.
[60] Y. Otsu and T. Shioya, *The Riemannian structure of Alexandrov spaces*, J. Differential Geom. **39** (1994), 629–658.
[61] Y. Otsu, K. Shiohama and T. Yamaguchi, *A new version of differentiable sphere theorem*, Invent. Math. **98** (1989), 219–228.
[62] G. Perelman, *Alexandrov's spaces with curvatures bounded from below II*, Preprint.
[63] G. Perelman, *Elements of Morse theory in Alexandrov spaces*, St. Petersburg Math. J. **5** (1994), 205–213.
[64] G. Perelman, *Widths of nonnegatively curved spaces*, Preprint.
[65] G. Perelman and A. Petrunin, *Quasigeodesics and gradient curves in Alexandrov spaces*, Workshop on Comparison Geometry, MSRI Publications (1993).
[66] G. Perelman and A. Petrunin, *Extremal subsets in Alexandrov spaces and the generalized Lieberman theorem*, St. Petersburg Math. J. **5** (1994), 215–227.
[67] S. Peters, *Convergence of Riemannian manifolds*, Comp. Math. **62** (1987), 3–16.
[68] P. Petersen, F. Wilhelm and S. Zhu, *Spaces on and beyond the boundary of existence*, J. Geom. Anal. **5** (1995), 419–426.
[69] A. Petrunin, *Quasigeodesics in multidimensional Alexandrov spaces*, Dissertation, Univ. Illinois (1995).
[70] A. Petrunin, *Applications of quasigeodesics and gradient curves*, Comparison Geometry (Berkeley, CA 1993–94), Math. Sci. Res. Inst. Pub., Vol. 30, Cambridge Univ. Press (1997), 203–219.

[71] A. Petrunin, *Parallel transportation for Alexandrov space with curvature bounded below*, Geom. Funct. Anal. **8** (1998), 123–148.
[72] C. Plaut, *Almost Riemannian spaces*, J. Differential Geom. **34** (1991), 515–537.
[73] C. Plaut, *A metric characterization of manifolds with boundary*, Comp. Math. **81** (1992), 337–354.
[74] C. Plaut, *Metric curvature, convergence, and topological finiteness*, Duke Math. J. **66** (1992), 43–57.
[75] C. Plaut, *Metric pinching of locally symmetric spaces*, Duke Math. J. **73** (1994), 155–162. *Correction*, Duke Math. J. **75** (1994), 527–528.
[76] C. Plaut, *Spaces of Wald–Berestovskii curvature bounded below*, J. Geom. Anal. **6** (1996), 113–134.
[77] C. Plaut, *Geometrizing locally compact infinite dimensional groups*, Trans. Amer. Math. Soc. **348** (1996), 941–962.
[78] W. Rinow, *Die innere Geometrie der metrischen Räume*, Springer, Berlin (1961).
[79] V.A. Sharafutdinov, *The radius of injectivity of a complete open manifold of nonnegative curvature*, Dokl. Akad. Nauk USSR **236** (1) (1976), 46–48.
[80] Z. Shen, *A regularity theorem for Alexandrov spaces*, Math. Nachr. **164** (1993), 91–102.
[81] Z. Shen, *Length and Finsler manifolds with curvature bounded from below*, Preprint.
[82] K. Shiohama, *An Introduction to the Geometry of Alexandrov Spaces*, Notes of the Series of Lecture held at the Seoul National University, No. 8 (1993).
[83] L.C. Siebenmann, *Deformation of homeomorphisms on stratified sets*, Comment. Math. Helv. **47** (1972), 123–163.
[84] N. Steenrod, *The Topology of Fiber Bundles*, Princeton Math. Series, Princeton University Press (1960).
[85] S. Shteingold, *Volume comparison for Alexandrov spaces*, Ind. Math. J. **43** (1994), 1349–1357.
[86] S. Shteingold, *Covering radii and paving diameters of Alexandrov spaces*, J. Geom. Anal. **8** (1998), 613–627.
[87] V.A. Toponogov, *Riemannian spaces of curvature bounded below*, Usp. Mat. Nauk **14** (1959), 87–130.
[88] A. Wald, *Begründung einer koordinatenlosen Differentialgeometrie der Flächen*, Ergebnisse eines mathematischen Kolloquiums **7** (1935), 24–46.
[89] F. Wilhelm, *On the filling radius of positively curved manifolds*, Invent. Math. **107** (1992), 653–668.
[90] F. Wilhelm, *Collapsing to almost Riemannian spaces*, Ind. Univ. Math J. **41** (1992), 1119–1142.
[91] T. Yamaguchi, *Collapsing and pinching under a lower curvature bound*, Ann. of Math. **133** (1991), 317–357.
[92] T. Yamaguchi, *A convergence theorem in the geometry of Alexandrov space*, Actes de la Ronde de Géométrie Différentielle (Luminy, 1992), Sémin. Cong. I, Soc. Math. France, Paris (1996), 601–642.

CHAPTER 17

Hyperbolic Manifolds

John G. Ratcliffe

Department of Mathematics, Vanderbilt University, Nashville, TN 37240, USA

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

From the point of view of geometric topology, the theory of closed hyperbolic manifolds represents the ideal situation in which homotopy theory reduces to geometry. Closed hyperbolic surfaces are classified up to homotopy type by the geometric invariants of area and orientability. Closed hyperbolic manifolds of dimension at least three satisfy the rigidity property that they are classified up to homotopy type by their isometry type with only finitely many manifolds having the same volume. Understanding to what extent the important properties of closed hyperbolic manifolds, such as rigidity, can be generalized to larger classes of manifolds and complexes has become a major theme in geometric topology. For this reason, every geometric topologist should be familiar with the basic theory of hyperbolic manifolds.

In this article, an overview of the theory of hyperbolic manifolds is presented with a primary emphasis on new developments in the theory of hyperbolic manifolds of dimension at least four. In order to place these developments in their proper context, a brief history of hyperbolic manifolds is given in Section 2 and a short survey of the theory of hyperbolic 3-manifolds is given in Section 3. New results on hyperbolic manifolds of finite volume of dimension at least four are discussed in Section 4 and the theory of geometrically finite hyperbolic manifolds of dimension at least four is outlined in Section 5.

2. History

When Riemann introduced the concept of an n-dimensional Riemannian manifold in his famous 1854 lecture [45] on the foundations of geometry, he gave as his principal examples the simply connected, complete, manifolds of constant sectional curvature. Riemann recognized the importance of these manifolds for the study of geometry because they are the simply connected Riemannian manifolds that possess free mobility between all coordinate frames. Riemann proved that a simply connected, complete, Riemannian n-manifold of constant sectional curvature is similar to either spherical n-space S^n, Euclidean n-space E^n, or hyperbolic n-space H^n according as its curvature is positive, zero, or negative, respectively. For this reason a Riemannian manifold of constant sectional curvature is called a *spherical, Euclidean,* or *hyperbolic* manifold according as its curvature is positive, zero, or negative, respectively. In order to simplify the discussion, we shall normalize the absolute value of the curvature of a spherical or hyperbolic manifold to be 1. We shall also assume that all spherical, Euclidean, or hyperbolic manifolds in this paper are connected.

In 1883 Klein [29] asserted his famous Uniformization Theorem that every closed Riemann surface is conformally equivalent to either a spherical, Euclidean, or hyperbolic surface. Klein's Uniformization Theorem is a beautiful connection between geometry and complex analysis that is central to the theory of Riemann surfaces. For a discussion, see Abikoff's article [1]. In 1890 Klein [30] proposed the problem of classifying all closed spherical, Euclidean, and hyperbolic surfaces. This classification was achieved by Fricke and Klein and was described in their two volume treatise on automorphic functions [17]. In particular, a closed surface supports a spherical, Euclidean, or hyperbolic structure according as its Euler characteristic is positive, zero, or negative, respectively.

An important feature of the classification of closed hyperbolic surfaces is that every closed surface with a negative Euler characteristic supports a continuum of distinct hyperbolic structures. Fricke and Klein proved that the space of hyperbolic structures on a closed orientable surface of genus $g > 1$ has dimension $6g - 6$. Understanding the space of hyperbolic structures on a surface has been one of the main themes in the theory of hyperbolic surfaces. For a discussion see Bers' article [5] and Thurston's article [51].

In 1891 Killing [28] recognized that a closed spherical, Euclidean, or hyperbolic n-manifold can be represented as an orbit space of a discrete group of isometries acting freely on S^n, E^n, or H^n, respectively. Killing proved that if n is even, then S^n and elliptic n-space P^n are the only closed spherical n-manifolds. Spherical n-manifolds, for odd n, which include lens spaces, provide a rich source of examples of manifolds whose classification up to homeomorphism, simply homotopy, and homotopy led to the development of much of geometric topology during the twentieth century. Spherical n-manifolds were classified up to isometry by Wolf in his 1967 treatise [58].

Every closed Euclidean n-manifold is finitely covered by a Euclidean n-torus, and so the classification of closed Euclidean n-manifolds is reduced to the classification of the finite groups of isometries of a Euclidean n-torus acting without fixed points. According to Bieberbach's solution of Hilbert's 18th problem, there are only finitely many closed Euclidean n-manifolds up to affine equivalence for each n. Bieberbach also proved that if two closed Euclidean n-manifolds have isomorphic fundamental groups, then they are affinely equivalent. Closed Euclidean 3- and 4-manifolds have been classified and a scheme for classifying closed Euclidean n-manifolds in general has been devised. For a discussion, see Charlap's text [14].

At the end of his 1891 paper [28], Killing proposed the problem of determining all discrete groups of isometries of hyperbolic n-space H^n that act freely on H^n. In 1926 Hopf [23] showed that a complete spherical, Euclidean, or hyperbolic n-manifold M can be represented as an orbit space of a discrete group Γ of isometries acting freely on S^n, E^n, or H^n, respectively, with Γ isomorphic to the fundamental group of M. Thus, Killing's problem is equivalent to determining all complete hyperbolic n-manifolds. In order to simplify the discussion, we shall now assume that a hyperbolic manifold is complete unless otherwise stated.

The most basic and important invariant of a hyperbolic manifold is its volume. Thus hyperbolic manifolds are divided naturally into two broad families, the manifolds of finite volume and the manifolds of infinite volume. The manifolds of finite volume are divided into two types, closed manifolds and open manifolds. The manifolds of infinite volume are divided into three types, the geometrically finite manifolds, the geometrically infinite manifolds with finitely generated fundamental group, and the manifolds with nonfinitely generated fundamental group.

The most important property of hyperbolic n-manifolds of finite volume for $n \geqslant 3$ is that they satisfy the following remarkable theorem which stands in direct contrast to the classification of closed hyperbolic surfaces.

THEOREM 2.1 (Rigidity theorem). *If two hyperbolic n-manifolds of finite volume, with $n \geqslant 3$, have isomorphic fundamental groups, then they are isometric to each other.*

The Rigidity theorem was proved by Mostow for closed manifolds in his 1973 study [37] and by Prasad for open manifolds in his 1973 paper [38]. The Rigidity theorem says that for $n \geqslant 3$ the isomorphism class of the fundamental group of a hyperbolic n-manifold of finite volume is a complete invariant of the isometry class of the manifold. From the point of view of geometric topology, the Rigidity Theorem says that for $n \geqslant 3$ two hyperbolic n-manifolds of finite volume are homotopically equivalent if and only if they are isometric. Thus the Rigidity theorem implies that for $n \geqslant 3$ all the geometric invariants of a hyperbolic n-manifold of finite volume are also homotopy invariants of the manifold.

3. Hyperbolic 3-manifolds

In 1979 William Thurston [49] revolutionized the theory of 3-manifolds when he proposed his Geometrization Conjecture which asserts that the interior of every compact 3-manifold can be canonically decomposed into pieces which admit one of eight different geometric structures. Of the eight geometries which occur, seven are well understood and the manifolds admitting them have been classified. For a discussion, see Scott's 1984 survey [47] and Thurston's 1997 text [52]. The last and most interesting geometry is that of hyperbolic 3-space. Hyperbolic geometry plays the primary role in the Geometrization Conjecture, since a generic closed irreducible 3-manifolds should have a hyperbolic structure. The Geometrization Conjecture has been proved by Thurston [50] for Haken 3-manifolds and many other non-Haken 3-manifolds. If the Geometrization Conjecture is true, then we will have a very satisfying classification of compact 3-manifolds.

A consequence of Thurston's theory is that the exterior of a smooth link L in the 3-sphere S^3 satisfies the Geometrization Conjecture. Now every closed orientable 3-manifold can be obtained by Dehn surgery on the exterior of some link $L \subset S^3$. One way of attacking the Geometrization Conjecture is to understand how the geometric structure on the exterior of a link $L \subset S^3$ changes under Dehn surgery. In particular, Thurston has proved that if $L \subset S^3$ is a link such that $S^3 - L$ has a hyperbolic structure, then almost all Dehn surgeries yield hyperbolic manifolds. For a survey on hyperbolic structures on knot complements, see Callahan and Reid's 1998 article [9].

The set of volumes of complete hyperbolic n-manifolds of finite volume is called the *volume spectrum* of hyperbolic n-manifolds. Using the technique of hyperbolic Dehn surgery, Thurston and Jørgensen have proved that the volume spectrum of hyperbolic 3-manifolds is a well-ordered subset of $\mathbb{R}$ and the set of manifolds with any given volume is finite. The order type of the volume spectrum of orientable hyperbolic 3-manifolds is ω^ω, which is the countable ordinal that describes the order type of polynomials in ω over $\mathbb{Z}$. This means that the volume spectrum of orientable 3-manifolds can be ordered as follows

$$v_1 < v_2 < v_3 < \cdots < v_\omega < v_{\omega+1} < \cdots < v_{2\omega} < \cdots < v_{\omega^2} < \cdots.$$

Here v_1 is the least volume of a closed orientable 3-manifold and the first limit point v_ω is the least volume of an orientable 3-manifold with one cusp. The first limit point of limit points v_{ω^2} is the least volume of an orientable 3-manifold with two cusps. For a discussion, see Gromov's article [19] and Thurston's article [49].

Although the volume function is finite-to-one on hyperbolic 3-manifolds, there is no bound on the number of manifolds with the same volume, since Wielenberg [57] has constructed arbitrarily large finite sets of nonisometric open hyperbolic 3-manifolds with the same finite volume and Apanasov and Gutsul [4] have constructed arbitrarily large finite sets of nonisometric closed hyperbolic 3-manifolds with the same volume.

Weeks [56] has written a computer program, called *SnapPea*, that will compute the hyperbolic invariants of a hyperbolic knot or link in S^3. The volumes of hyperbolic knots and links in S^3, with small crossing number, have been tabulated by Adams, Hildebrand and Weeks [2] using SnapPea. The program SnapPea will also compute the hyperbolic invariants of a closed hyperbolic 3-manifold obtained by Dehn surgery on a hyperbolic knot or link in S^3. Closed orientable hyperbolic 3-manifolds of small complexity have been enumerated by Matveev and Fomenko [33] using the technique of hyperbolic Dehn surgery.

Enumerating hyperbolic 3-manifolds of finite volume in order of increasing volume seems to be difficult. To begin with, the smallest volume of a closed orientable manifold v_1 is unknown. From the investigations of Weeks and of Matveev and Fomenko, the smallest known volume of a closed, orientable, hyperbolic 3-manifold is about 0.9427. For further discussion, see Matveev and Fomenko's article [33] and Meyerhoff's article [34].

The volume spectrum of open hyperbolic 3-manifolds is better understood. Cao and Meyerhoff [13] have determined the open orientable hyperbolic 3-manifolds of smallest volume and proved that their volume v_ω is 2.0298832... which is the volume of the figure-eight knot complement.

In view of the Geometrization Conjecture, it is important to recognize when a closed 3-manifold M admits a hyperbolic structure. A necessary condition is that the fundamental group of M is isomorphic to the fundamental group of a closed hyperbolic 3-manifold. Gromov [18] and Cannon and Cooper [12] have proved that a group G is isomorphic to the fundamental group of a closed hyperbolic 3-manifold if and only if G is finitely generated, torsion-free, and the associated Cayley graph is quasi-isometric to hyperbolic 3-space H^3. If the Cayley graph of a group G is quasi-isometric to H^3, then the group G is *word hyperbolic* in the sense of Gromov [20]. Cannon [11] has asked whether every closed 3-manifold with a word hyperbolic fundamental group admits a hyperbolic structure.

A hyperbolic 3-manifold $M = H^3/\Gamma$ is said to be *geometrically finite* if its associated Kleinian group Γ has a finite-sided Dirichlet fundamental polyhedron. Every hyperbolic 3-manifold of finite volume is geometrically finite. The ends of a geometrically finite hyperbolic 3-manifold have simple geometry. Let $O(\Gamma)$ be the *ordinary set* of Γ on the sphere at infinity of H^3. Then $\overline{M} = (H^3 \cup O(\Gamma))/\Gamma$ is a 3-manifold-with-boundary. The boundary of $\overline{M}$ is a disjoint union of Riemann surfaces, since Γ acts conformally on $O(\Gamma)$. Marden [32] has proved that M is geometrically finite if and only if $\overline{M}$ is the disjoint union of a compact submanifold and a finite union of rank one and rank two cusps. Geometrically finite hyperbolic 3-manifolds are well understood.

If M is geometrically finite, then the fundamental group of M is finitely generated. There are hyperbolic 3-manifolds with finitely generated fundamental group that are not geometrically finite. Let $M = H^3/\Gamma$ be a hyperbolic 3-manifold with Γ finitely generated. Ahlfors [3] has proved that the boundary of $\overline{M}$ is a finite union of Riemann surfaces of finite type. Ahlfors' finiteness theorem suggests that a hyperbolic 3-manifold with a finitely generated fundamental group is not too far from being geometrically finite. If $O(\Gamma)$ is nonempty, then the hyperbolic structure on M can be deformed into other hyperbolic structures. Understanding the deformations of the hyperbolic structure of a hyperbolic 3-manifold with a finitely generated fundamental group is one of the main themes in the theory of Kleinian groups. For a discussion, see Bers' paper [5], Thurston's paper [49], and Minsky's paper [36].

An end of a hyperbolic 3-manifold is said to be *geometrically finite* if it has the same geometry as an end of a geometrically finite hyperbolic 3-manifold. A non-cusp geometrically finite end of a hyperbolic 3-manifold is exponentially expanding. Thurston [48] has proposed another possible model for the geometry of a non-cusp end of a hyperbolic 3-manifold. An end of a hyperbolic 3-manifold is said to be *simply degenerate* if it has a neighborhood U homeomorphic to $S \times [0, \infty)$ and there is a sequence of singular surfaces $\{f_n : S \to U\}$ such that $f_n(S)$ is homotopic to the compact surface $S \times \{0\}$ within U for each n, the intrinsic geometry of $f_n(S)$ has curvature $\leqslant -1$ for each n, and the sequence $\{f_n(S)\}$ eventually leaves any compact subset of U. A hyperbolic 3-manifold M with a finitely generated fundamental group is said to be *geometrically tame* if all its ends are geometrically finite or simply degenerate. Thurston [49] has asked whether every hyperbolic 3-manifold with a finitely generated fundamental group is geometrically tame. Bonahon [6] has proved that if the fundamental group of M is freely indecomposable, then M is geometrically tame. Canary [10] has proved that M is geometrically tame if and only if M is homeomorphic to the interior of a compact 3-manifold.

The ends of a geometrically tame hyperbolic 3-manifold can be approximately characterized by their *end invariants*. The invariant of a non-cusp geometrically finite end of M is the Riemann surface boundary component of $\overline{M}$ corresponding to the end of M. The invariant of a simply degenerate end is a geodesic lamination on the surface S determined by the end. Thurston [49] has asked whether the isometry type of a hyperbolic 3-manifold with a finitely generated fundamental group is determined by its topological type and its list of end invariants. For a discussion, see Minsky's paper [36].

4. Hyperbolic n-manifolds of finite volume with $n \geqslant 4$

The volume spectrum of hyperbolic n-manifolds, for $n \geqslant 4$, is much simpler than the volume spectrum of hyperbolic 3-manifolds because of the following theorem of Wang [55].

THEOREM 4.1 (Wang's Finiteness theorem). *If $n \geqslant 4$, then for each real number r there are only finitely many isometry classes of n-dimensional hyperbolic manifolds M with* $\mathrm{Vol}(M) \leqslant r$.

Wang's Finiteness theorem implies that the volume spectrum of hyperbolic n-manifolds, for $n \geqslant 4$, is a closed discrete subset of $\mathbb{R}$ with only finitely many manifolds having the same volume. Millson [35] has proved that there is a closed hyperbolic n-manifold with a positive first Betti number for each dimension n. Such a manifold has an m-fold covering for each positive integer m. Therefore the volume spectrum of hyperbolic n-manifolds is infinite for each n.

If M is a hyperbolic n-manifold of finite volume, with n even, then the Gauss–Bonnet theorem [24,27] says that

$$\mathrm{Vol}(M) = (-1)^{n/2}\mathrm{Vol}(P^n)\chi(M).$$

The volume of elliptic n-space P^n is half the volume of the unit n-sphere S^n. Thus the volume spectrum of hyperbolic n-manifolds, with n even, is an infinite subset of the positive integral multiples of $\mathrm{Vol}(P^n)$ where

$$\mathrm{Vol}(P^n) = \frac{(2\pi)^{n/2}}{(n-1)(n-3)\cdots 3\cdot 1}.$$

If M is a hyperbolic 4-manifold of finite volume, then the Gauss–Bonnet theorem says that

$$\mathrm{Vol}(M) = \frac{4}{3}\pi^2\chi(M).$$

The next theorem was proved by Ratcliffe and Tschantz [44].

THEOREM 4.2. *The volume spectrum of hyperbolic* 4*-manifolds is the set of all positive integral multiples of* $4\pi^2/3$.

The theorem is proved by constructing a hyperbolic 4-manifold M_1 with finite volume, Euler characteristic one, and a positive first Betti number. The manifold M_1 has minimum volume by the Gauss–Bonnet theorem, since $\chi(M_1) = 1$. Moreover, M_1 has an m-fold covering for each positive integer m, since its first Betti number is positive. Hence there are hyperbolic 4-manifolds of every possible finite volume allowed by the Gauss–Bonnet theorem.

The manifold M_1 is constructed by gluing together the sides of a regular ideal 24-cell P in hyperbolic 4-space so as to satisfy the conditions of Poincaré's fundamental polyhedron theorem [39]. Therefore M_1 is an open manifold with cusps. The regular polytope P has 24 sides each of which is a regular ideal octahedron. The volume of P is $4\pi^2/3$. In the conformal ball model of hyperbolic 4-space,

$$B^4 = \{x \in E^4\colon |x| < 1\},$$

the polytope P has vertices $(\pm 1,0,0,0)$, $(0,\pm 1,0,0)$, $(0,0,\pm 1,0)$, $(0,0,0,\pm 1)$, and $(\pm\frac{1}{2},\pm\frac{1}{2},\pm\frac{1}{2},\pm\frac{1}{2})$. Each side-pairing transformation of P that determines the gluing of M_1 is the composite of a reflection in a coordinate plane followed by the reflection of P in one of its sides. Therefore the manifold M_1 is orientable.

The manifold M_1 is an arithmetic manifold with respect to the hyperboloid model of hyperbolic 4-space in Lorentzian 5-space. Recall that *Lorentzian* $(n+1)$*-space* is the inner product space $\mathbb{R}^{n,1}$ consisting of real $(n+1)$-space $\mathbb{R}^{n+1}$ with the *Lorentzian inner product*

$$x \circ y = x_1y_1 + x_2y_2 + \cdots + x_ny_n - x_{n+1}y_{n+1}.$$

A *Lorentz transformation* of $\mathbb{R}^{n,1}$ is a linear transformation of $\mathbb{R}^{n+1}$ that preserves Lorentzian inner products. The matrix of a Lorentz transformation with respect to the standard basis of $\mathbb{R}^{n+1}$ is called a *Lorentzian matrix*.

The *hyperboloid model* of hyperbolic n-space is the metric space

$$H^n = \{x \in \mathbb{R}^{n,1}\colon\ x \circ x = -1 \text{ and } x_{n+1} > 0\}$$

with metric defined by

$$\cosh d(x, y) = -x \circ y.$$

A Lorentz transformation of $\mathbb{R}^{n,1}$ is either *positive* or *negative* according as it maps H^n to H^n or $-H^n$. The isometries of H^n correspond, by restriction, to the positive Lorentz transformations of $\mathbb{R}^{n,1}$. A Lorentzian matrix is said to be *positive* if the corresponding Lorentz transformation is positive.

Let Γ^n be the group of positive Lorentzian $(n + 1) \times (n + 1)$ matrices with integer entries. The group Γ^n is an infinite, discrete, arithmetic subgroup of the group $\mathrm{O}(n, 1)$ of Lorentzian $(n + 1) \times (n + 1)$ matrices. The *principal congruence two subgroup* of Γ^n is the group Γ_2^n of all matrices in Γ^n that are congruent to the identity matrix modulo two. The congruence two subgroup Γ_2^n is not torsion-free, but it only has 2-torsion.

The side-pairing transformations of the regular ideal 24-cell P that determine the gluing of the hyperbolic 4-manifold M_1 correspond, via stereographic projection from B^4 to H^4, to integral Lorentzian 5×5 matrices which lie in the congruence two subgroup Γ_2^4. Thus M_1 is an arithmetic hyperbolic 4-manifold.

By Wang's finiteness theorem, there are only finitely many hyperbolic 4-manifolds, up to isometry, of minimum volume. We have also proved the following classification theorem.

THEOREM 4.3. *There are, up to isometry,* 1171 *hyperbolic* 4*-manifolds of minimum volume whose fundamental group is a subgroup of the congruence two group of integral Lorentzian* 5×5 *matrices. Only* 22 *of these manifolds are orientable.*

All 1171 of these hyperbolic 4-manifolds are obtained by gluing together the sides of a regular ideal 24-cell P. They are all open manifolds with five or six cusps. They are classified by considering the self-intersection set of the boundary of a maximal cusp, and then applying a cut and paste argument. Table 1 lists invariants of the 22 orientable manifolds. The column headed by H_i lists the ith homology groups. In particular, the first 15 manifolds in Table 1 have a positive first Betti number. The six closed, orientable, Euclidean 3-manifolds $A, B, \ldots, F$ are in the same order as in Hantzsche and Wendt's paper [22] with A the 3-torus and F the Hantzsche–Wendt 3-manifold.

A hyperbolic 4-manifold that can be obtained by gluing together the sides of a regular ideal 24-cell is called a *24-cell manifold*. A 24-cell manifold has minimum volume. A partial computer search for 24-cell manifolds has quickly turned up hundreds of orientable 24-cell manifolds which suggests that there are probably millions of 24-cell manifolds. A complete search and classification of all 24-cell manifolds is not possible with present computer technology.

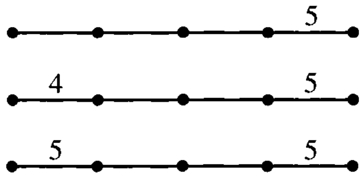
Table 1
Invariants of the orientable, congruence two, 24-cell manifolds

Number	Symmetries	H_1	H_2	H_3	Cusp Types
1	16	$\mathbb{Z}^3 \oplus \mathbb{Z}_2^3$	$\mathbb{Z}^7$	$\mathbb{Z}^4$	*AAABF*
2	16	$\mathbb{Z}^2 \oplus \mathbb{Z}_2^4$	$\mathbb{Z}^6$	$\mathbb{Z}^4$	*AABBF*
3	16	$\mathbb{Z}^2 \oplus \mathbb{Z}_2^4$	$\mathbb{Z}^6$	$\mathbb{Z}^4$	*AABBF*
4	16	$\mathbb{Z}^2 \oplus \mathbb{Z}_2^4$	$\mathbb{Z}^6$	$\mathbb{Z}^4$	*AABBF*
5	16	$\mathbb{Z}^2 \oplus \mathbb{Z}_2^4$	$\mathbb{Z}^6$	$\mathbb{Z}^4$	*AABBF*
6	16	$\mathbb{Z}^2 \oplus \mathbb{Z}_2^4$	$\mathbb{Z}^6$	$\mathbb{Z}^4$	*ABBBF*
7	48	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
8	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
9	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
10	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
11	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
12	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
13	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*ABBBF*
14	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*BBBBF*
15	16	$\mathbb{Z} \oplus \mathbb{Z}_2^5$	$\mathbb{Z}^5$	$\mathbb{Z}^4$	*BBBBF*
16	16	$\mathbb{Z}_2^6$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*BBBBF*
17	16	$\mathbb{Z}_2^6$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*BBBBF*
18	16	$\mathbb{Z}_2^6$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*BBBBF*
19	16	$\mathbb{Z}_2^6$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*BBFFF*
20	16	$\mathbb{Z}_2^5 \oplus \mathbb{Z}_4$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*ABFFF*
21	16	$\mathbb{Z}_2^5 \oplus \mathbb{Z}_4$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*ABFFF*
22	16	$\mathbb{Z}_2^5 \oplus \mathbb{Z}_4$	$\mathbb{Z}^4$	$\mathbb{Z}^4$	*ABFFF*

The group of symmetries of the regular tessellation of H^4 obtained by reflecting in the sides of the regular ideal 24-cell P is a hyperbolic reflection group Γ with respect to a Coxeter simplex Δ whose Coxeter graph is

The Coxeter simplex Δ is the unique noncompact, hyperbolic, Coxeter orthoscheme in dimension 4. The 24-cell P is subdivided into 1152 copies of Δ. The groups of the 1171 congruence two 24-cell manifolds are torsion-free subgroups of index 1152 of the reflection group Γ.

There are exactly three compact, hyperbolic, Coxeter orthoschemes in dimension 4. We denote these three orthoschemes by Δ_3, Δ_4, and Δ_5. Their Coxeter graphs are



The hyperbolic reflection groups Γ_k with respect to Δ_k is the group of symmetries of the regular tessellations of H^4 obtained by reflecting in the sides of a regular 120-cell P_k in H^4 with dihedral angle $2\pi/k$ for $k = 3, 4, 5$. The 120-cell P_k is subdivided into 14400 copies of Δ_k, for $k = 3, 4, 5$.

Davis [16] has constructed a closed, orientable, hyperbolic 4-manifold M_5 by gluing together the opposite sides of P_5. The Euler characteristic of M_5 is 26, since the Euler characteristic of Γ_5 is 26/14400. Ratcliffe and Tschantz have constructed a closed, nonorientable, hyperbolic 4-manifold M_4 by gluing together two copies of P_4. The Euler characteristic of M_4 is 17, since the Euler characteristic of Γ_4 is 17/28800. The Euler characteristic of Γ_3 is 1/14400, and so it may be possible to construct a closed hyperbolic 4-manifold M_3 with Euler characteristic one and minimum volume by gluing together the sides of P_3. The manifold M_3 would have to be nonorientable because of the next theorem.

THEOREM 4.4. *Every closed, orientable, hyperbolic manifold has an even Euler characteristic.*

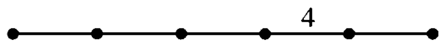
PROOF. Let M be a closed, orientable, hyperbolic n-manifold. Every closed orientable n-manifold has an even Euler characteristic when n is not a multiple of four, and so we may assume that n is a multiple of four. By a theorem of Chern [15], all the Pontrjagin numbers of M are zero. Therefore, the signature $\sigma(M)$ of M is zero by the signature theorem, and so we have

$$\chi(M) \equiv \beta_{n/2}(M) \equiv \sigma(M) \equiv 0 \bmod 2. \qquad \square$$

There is not much known about closed hyperbolic 4-manifolds. In particular, the volume spectrum of closed orientable hyperbolic 4-manifolds is unknown. The closed, orientable, hyperbolic 4-manifold of smallest known volume is the Davis manifold M_5. Hyperbolic 4-manifolds of small volume have become of interest in Cosmology. For a discussion, see Ratcliffe and Tschantz's 1998 article [43].

By generalizing the construction of the congruence two 24-cell hyperbolic 4-manifolds, Ratcliffe and Tschantz have constructed an open, nonorientable, arithmetic, hyperbolic 5-manifold M_2, with a positive first Betti number, whose volume is $28\zeta(3)$ where ζ is the Riemann zeta function. The manifold M_2 is constructed by gluing together the sides of a noncompact polytope Q in hyperbolic 5-space so as to satisfy the conditions of Poincaré's fundamental polyhedron theorem. The manifold M_2 is arithmetic because the side-pairing transformations of Q that determine the gluing of M_2 correspond to integral Lorentzian 6×6 matrices which lie in the congruence two subgroup Γ_2^5.

The volume of the manifold M_2 is the volume of Q. The polytope Q can be subdivided into 184320 copies of the Coxeter orthoscheme whose Coxeter graph is



The volume of this orthoscheme was computed to be $7\zeta(3)/46080$ by Kellerhals [26]. Therefore the volume of M_2 is $28\zeta(3)$. It is worth noting that Apéry [53] has proved that $\zeta(3)$ is irrational.

Let Γ^n be the group of positive Lorentzian $(n+1) \times (n+1)$ matrices with integer entries. For k a positive integer, the *principal congruence subgroup* of Γ^n of *level* k is the group Γ_k^n of all matrices in Γ^n that are congruent to the identity matrix modulo k.

The group Γ_k^n is torsion-free if and only if $k > 2$. Therefore Γ_k^n acts freely on H^n if and only if $k > 2$. Define $M_k^n = H^n/\Gamma_k^n$. Then M_k^n is a hyperbolic n-orbifold for $k = 1, 2$, and a hyperbolic n-manifold for $k > 2$. The manifolds M_k^n, with $k > 2$, are the simplest to define, arithmetic, hyperbolic n-manifolds.

Note that $\Gamma_1^n = \Gamma^n$, and so $M_1^n = H^n/\Gamma^n$. Ratcliffe and Tschantz [42] have computed the volume of M_1^n using Siegel's analytic theory of quadratic forms [46]. Let B_k be the kth Bernoulli number with even index notation so that

$$B_2 = \frac{1}{6}, \qquad B_4 = -\frac{1}{30}, \qquad B_6 = \frac{1}{42}, \ \ldots .$$

If n is even, then

$$\mathrm{Vol}\big(M_1^n\big) = \big(2^{n/2} \pm 1\big) \prod_{k=1}^{n/2} (-1)^{k-1} B_{2k} \cdot \frac{\pi^{n/2}}{n!}$$

with the plus sign if $n \equiv 0, 2 \bmod 8$ and the minus sign if $n \equiv 4, 6 \bmod 8$.

Let χ be the nontrivial Dirichlet character modulo 4 and consider the Dirichlet L-function

$$L(s) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \cdots .$$

If $n \equiv 3 \bmod 4$, then

$$\mathrm{Vol}\big(M_1^n\big) = \prod_{k=1}^{(n-1)/2} (-1)^{k-1} \frac{B_{2k}}{2k} \cdot L\big((n+1)/2\big).$$

Consider the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

If $n \equiv 1 \bmod 4$, then

$$\mathrm{Vol}(M_1^n) = \left(2^{(n+1)/2} - 1\right)\left(2^{(n-1)/2} \pm 1\right) \prod_{k=1}^{(n-1)/2} (-1)^{k-1} \frac{B_{2k}}{8k} \cdot \frac{\zeta((n+1)/2)}{2}$$

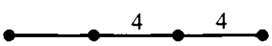
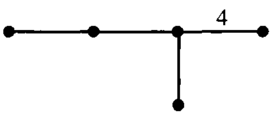
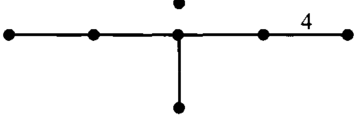
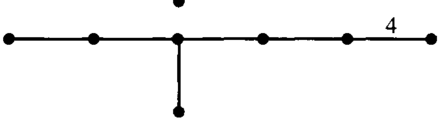
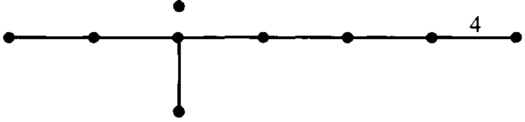
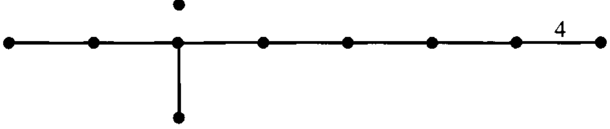
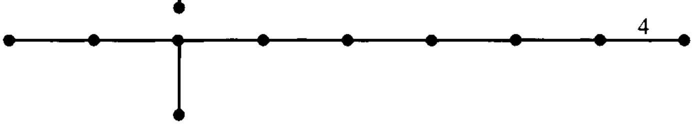
with the plus sign if $n \equiv 1 \bmod 8$ and the minus sign if $n \equiv 5 \bmod 8$.

The group Γ^1 consists of the two diagonal matrices $\mathrm{diag}(\pm 1, 1)$, and so M_1^1 is a hyperbolic ray and $\mathrm{Vol}(M_1^1) = \infty$. The group Γ^n, for $n = 2, 3, \ldots, 9$, is a discrete, hyperbolic, reflection group with M_1^n isometric to a Coxeter n-simplex in H^n. Table 2 lists the Coxeter diagrams and volumes of M_1^n for $n = 2, 3, \ldots, 9$.

The group Γ_k^n is a subgroup of finite index of Γ^n. Therefore, we have that

$$\mathrm{Vol}(M_k^n) = \left[\Gamma^n : \Gamma_k^n\right] \mathrm{Vol}(M_1^n).$$

Table 2
Table of Coxeter diagrams and volumes of M_1^n for $n = 2, 3, \ldots, 9$

M_1^2		$\frac{\pi}{4}$
M_1^3		$\frac{L(2)}{12}$
M_1^4		$\frac{\pi^2}{1440}$
M_1^5		$\frac{7\zeta(3)}{15360}$
M_1^6		$\frac{\pi^3}{777600}$
M_1^7		$\frac{L(4)}{362880}$
M_1^8		$\frac{17\pi^4}{9144576000}$
M_1^9		$\frac{527\zeta(5)}{44590694400}$

Let p be an odd prime number, let $n > 1$, and let $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z}$. Ratcliffe and Tschantz [42] have proved that

$$[\Gamma^n : \Gamma^n_p] = \begin{cases} \frac{1}{2}|\mathrm{O}(n,1;\mathbb{Z}_p)|, & \text{if } p \equiv 1,7 \bmod 8; \\ |\mathrm{O}(n,1;\mathbb{Z}_p)|, & \text{if } p \equiv 3,5 \bmod 8. \end{cases}$$

The order of the group $\mathrm{O}(n,1;\mathbb{Z}_p)$ is given by the classical formula

$$|\mathrm{O}(n,1;\mathbb{Z}_p)| = \begin{cases} 2\prod_{k=1}^{n/2} p^{2k-1}(p^{2k}-1), & \text{if } n \text{ is even;} \\ 2\left(p^{(n+1)/2} - \left(\frac{(-1)^{(n-1)/2}}{p}\right)\right)\prod_{k=1}^{(n-1)/2} p^{2k}(p^{2k}-1), & \text{if } n \text{ is odd.} \end{cases}$$

Let $a = 0$ or 1 according as $p \equiv 1,7 \bmod 8$ or $p \equiv 3,5 \bmod 8$.

If n is even and p is an odd prime number, then we have that

$$\mathrm{Vol}(M^n_p) = 2^a(2^{n/2} \pm 1)\prod_{k=1}^{n/2} p^{2k-1}(p^{2k}-1)|B_{2k}| \cdot \frac{\pi^{n/2}}{n!}$$

with the plus sign if $n \equiv 0, 2 \bmod 8$ and the minus sign if $n \equiv 4, 6 \bmod 8$.

If $n \equiv 3 \bmod 4$ and p is an odd prime number, then we have that

$$\mathrm{Vol}(M^n_p) = 2^a(p^{(n+1)/2} \pm 1)\prod_{k=1}^{(n-1)/2} p^{2k}(p^{2k}-1)\frac{|B_{2k}|}{2k} \cdot L((n+1)/2)$$

with the plus sign if $p \equiv 3 \bmod 4$ and the minus sign if $p \equiv 1 \bmod 4$.

If $n \equiv 1 \bmod 4$ and p is an odd prime number, then we have that

$$\begin{aligned} \mathrm{Vol}(M^n_p) &= 2^a(2^{(n+1)/2}-1)(2^{(n-1)/2} \pm 1)(p^{(n+1)/2}-1) \\ &\quad \cdot \prod_{k=1}^{(n-1)/2} p^{2k}(p^{2k}-1)\frac{|B_{2k}|}{8k} \cdot \frac{\zeta((n+1)/2)}{2} \end{aligned}$$

with the plus sign if $n \equiv 1 \bmod 8$ and the minus sign if $n \equiv 5 \bmod 8$.

The volumes of M^5_p for the first ten odd primes are given below:

$$\begin{aligned} \mathrm{Vol}(M^5_3) &= 22113\zeta(3)/2, \\ \mathrm{Vol}(M^5_5) &= 26446875\zeta(3), \\ \mathrm{Vol}(M^5_7) &= 2112387795\zeta(3), \\ \mathrm{Vol}(M^5_{11}) &= 7545654056325\zeta(3)/2, \end{aligned}$$

$$\begin{aligned}
\mathrm{Vol}(M_{13}^5) &= 46355018040513\zeta(3),\\
\mathrm{Vol}(M_{17}^5) &= 1299695706128736\zeta(3),\\
\mathrm{Vol}(M_{19}^5) &= 27593035151946705\zeta(3)/2,\\
\mathrm{Vol}(M_{23}^5) &= 121273347969963651\zeta(3),\\
\mathrm{Vol}(M_{29}^5) &= 7855460647805208165\zeta(3),\\
\mathrm{Vol}(M_{31}^5) &= 10682306350810239600\zeta(3).
\end{aligned}$$

These volumes are obviously much larger than the volume $28\zeta(3)$ of the congruence two hyperbolic 5-manifold mentioned earlier.

From the Gauss–Bonnet theorem, we can derive explicit formulas for the Euler characteristic of the manifold M_p^n for each odd prime number p. If n is odd and $n > 1$, then $\chi(M_p^n) = 0$. If n is even, then the Euler characteristic of M_p^n is given by the formula

$$\chi(M_p^n) = (-1)^{n/2}2^a(2^{n/2} \pm 1)\prod_{k=1}^{n/2} p^{2k-1}(p^{2k}-1)\frac{|B_{2k}|}{4k},$$

where $a = 0$ or 1 according as $p \equiv 1, 7 \bmod 8$ or $p \equiv 3, 5 \bmod 8$ and with the plus sign if $n \equiv 0, 2 \bmod 8$ and the minus sign if $n \equiv 4, 6 \bmod 8$.

The values of $\chi(M_p^4)$ for the first ten odd primes are given below:

$$\begin{aligned}
\chi(M_3^4) &= 54,\\
\chi(M_5^4) &= 9750,\\
\chi(M_7^4) &= 144060,\\
\chi(M_{11}^4) &= 26793030,\\
\chi(M_{13}^4) &= 142747878,\\
\chi(M_{17}^4) &= 1046351088,\\
\chi(M_{19}^4) &= 6368787270,\\
\chi(M_{23}^4) &= 21535443996,\\
\chi(M_{29}^4) &= 437714992470,\\
\chi(M_{31}^4) &= 426445056960.
\end{aligned}$$

The values of $\chi(M_p^6)$ for the first ten odd primes are given below:

$$\begin{aligned}
\chi(M_3^6) &= -44226,\\
\chi(M_5^6) &= -2203906250,\\
\chi(M_7^6) &= -1318755876760,\\
\chi(M_{11}^6) &= -35390554791603550,
\end{aligned}$$

$$\chi(M^6_{13}) = -1184384445755267302,$$
$$\chi(M^6_{17}) = -166020530109472479168,$$
$$\chi(M^6_{19}) = -343472991750108 1492650,$$
$$\chi(M^6_{23}) = -9499620852679716889 5304,$$
$$\chi(M^6_{29}) = -247238242187409767355743 50,$$
$$\chi(M^6_{31}) = -501635155793770474356608 00.$$

The manifolds M^n_p, with $n > 1$ and p an odd prime, have distinct volumes and so they are mutually nonisometric. From the volume formulas for M^n_p, we see that these manifolds are large; nevertheless, the manifolds M^n_p are beautiful because they have large symmetry groups.

All of the above hyperbolic n-manifolds, with $n \geqslant 4$, are arithmetic. There are three ways of constructing arithmetic, hyperbolic, discrete groups. For a discussion, see Vinberg and Shvartsman [54]. From these groups, arithmetic hyperbolic manifolds can be constructed as orbit spaces of torsion-free congruence subgroups. For a discussion, see Gromov and Piatetski-Shapiro [21] and Li and Millson [31]. All arithmetic hyperbolic manifolds have finite volume. The properties of arithmetic hyperbolic n-manifolds for $n \geqslant 4$ need to be more fully investigated.

The only known examples of nonarithmetic hyperbolic n-manifolds of finite volume, with $n \geqslant 6$, are those constructed by Gromov and Piatetski-Shapiro [21] by cutting and pasting arithmetic hyperbolic manifolds along totally geodesic hypersurfaces. For a discussion, see Vinberg and Shvartsman [54]. Gromov and Piatetski-Shapiro [21] have asked whether every hyperbolic n-manifold of finite volume, with $n \geqslant 4$, can be obtained by cutting and pasting arithmetic hyperbolic manifolds along totally geodesic hypersurfaces.

5. Geometrically finite hyperbolic n-manifolds with $n \geqslant 4$

A discrete group Γ of isometries of H^3 is said to be *geometrically finite* if Γ has a finite-sided Dirichlet fundamental polyhedron. The obvious generalization of this definition to higher dimensions is too restrictive because there are geometrically finite discrete groups of isometries of H^n, with $n \geqslant 4$, all of whose Dirichlet fundamental polyhedra are infinite-sided. In order to define a geometrically finite discrete group, we pass to the *conformal ball model* of hyperbolic n-space

$$B^n = \{x \in E^n \colon |x| < 1\}.$$

The sphere at infinity of B^n is S^{n-1}. Every isometry of B^n extends to a Möbius transformation of $E^n \cup \{\infty\}$. Let an overline denote the closure of a subset of $E^n \cup \{\infty\}$.

DEFINITION 5.1. A convex polyhedron P in B^n is *geometrically finite* if and only if for each point x of $\overline{P} \cap S^{n-1}$ there is an open neighborhood N of x in E^n that meets only the sides of P incident with x.

For example, every finite-sided convex polyhedron in B^n is geometrically finite.

DEFINITION 5.2. A *cusp point* of a convex polyhedron P in B^n is a point c of $\overline{P} \cap S^{n-1}$ for which there is an open neighborhood N of c in E^n such that the intersection of the Euclidean closures of all the sides of P that meet N is c.

Although a geometrically finite polyhedron P may have an infinite number of sides, P has only finitely many cusp points and all but finitely many sides of P are incident with a cusp point of P.

Bowditch [7] has proposed five different equivalent definitions of a geometrically finite discrete group Γ of isometries of B^n for $n \geqslant 4$, but the following new definition seems to be more natural.

DEFINITION 5.3. A discrete group Γ of isometries of B^n is *geometrically finite* if and only if Γ has a geometrically finite, Dirichlet, fundamental polyhedron.

Bowditch [7] has proved that a geometrically finite group Γ has an infinite-sided Dirichlet domain if and only if Γ has a parabolic element with an infinite order twist. Consequently if $n = 3$, every Dirichlet domain of a geometrically finite group Γ is finite-sided. Thus our definition of a geometrically finite group is equivalent to the standard definition when $n = 3$.

Before we can further discuss geometrically finite groups, we need to recall some of the basic definitions of limit points of groups of isometries of B^n.

DEFINITION 5.4. A point a of S^{n-1} is a *limit point* of a discrete group Γ of isometries of B^n if there is a point x of B^n and a sequence $\{g_i\}_{i=1}^{\infty}$ of elements of Γ such that $\{g_i x\}_{i=1}^{\infty}$ converges to a. The *limit set* of Γ is the set $L(\Gamma)$ of all limit points of Γ.

For example, the fixed points of either a parabolic or hyperbolic element of Γ are limit points of Γ.

DEFINITION 5.5. A discrete group Γ of isometries of B^n is *elementary* if the limit set of Γ is finite.

If Γ is elementary, then $L(\Gamma)$ consists of either 0, 1, or 2 points and Γ is said to be of *elliptic*, *parabolic*, or *hyperbolic type*, respectively. An elementary group Γ is of elliptic type if and only if Γ is finite, of parabolic type if and only if Γ has an infinite free abelian subgroup of finite index generated by parabolic elements, and of hyperbolic type if and only if Γ has an infinite cyclic subgroup of finite index generated by a hyperbolic element.

DEFINITION 5.6. A point a of S^{n-1} is a *conical limit point* of a discrete group Γ of isometries of B^n if there is a point x of B^n, a sequence $\{g_i\}_{i=1}^{\infty}$ of elements of Γ, a hyperbolic ray R in B^n ending at a, and an $r > 0$ such that $\{g_i x\}_{i=1}^{\infty}$ converges to a within the r-neighborhood $N(R, r)$ of R in B^n.

For example, a point a of S^{n-1} fixed by a hyperbolic element h of a discrete group Γ of isometries of B^n is a conical limit point of Γ. A conical limit point of Γ has the geometric property that it cannot lie on the Euclidean boundary of any convex fundamental polyhedron for Γ.

We now pass to the *upper half-space model* hyperbolic n-space

$$U^n = \{x \in E^n \colon x_n > 0\}.$$

Let Γ be a discrete group of isometries of U^n such that ∞ is fixed by a parabolic element of Γ. Then the stabilizer Γ_∞ is an elementary group of parabolic type. Therefore Γ_∞ corresponds under Poincaré extension to a discrete group of isometries of E^{n-1}. Hence, there is a Γ_∞-invariant m-plane Q of E^{n-1} such that Q/Γ_∞ is compact. Let $r > 0$ and let $N(Q,r)$ be the r-neighborhood of Q in E^n. Then $N(Q,r)$ is invariant under Γ_∞. Now set

$$U(Q,r) = \overline{U}^n - \overline{N}(Q,r).$$

Then $U(Q,r)$ is an open Γ_∞-invariant subset of $\overline{U}^n$. Note that if $m = n-1$, then $U(Q,r)$ is a horoball based at ∞. If $m < n-1$, then $U(Q,r)$ is the complement in $\overline{U}^n$ of the tunnel $N(Q,r) \cap \overline{U}^n$. The set $U(Q,r)$ is said to be a *cusped region* for Γ based at ∞ if and only if for all g in $\Gamma - \Gamma_\infty$, we have

$$U(Q,r) \cap gU(Q,r) = \emptyset.$$

Let c be a point of E^{n-1} fixed by a parabolic element of a discrete group Γ of isometries of U^n. A subset U of $\overline{U}^n$ is a *cusped region* for Γ based at c if and only if and only if upon conjugating Γ so that $c = \infty$, the set U transforms to a cusped region for Γ based at ∞.

DEFINITION 5.7. A *cusped limit point* of a discrete group Γ of isometries of U^n is a fixed point c of a parabolic element of Γ such that there is a cusped region U for Γ based at c.

A cusped limit point c of Γ has the geometric property that if P is a convex fundamental polyhedron for Γ, then there is an element g of Γ such that c is a cusp point of gP. In particular, no conical limit point is a cusped limit point. The following theorem is proved in Chapter 12 of [39].

THEOREM 5.1. *Let Γ be a discrete group of isometries of B^n. Then the following are equivalent*:

(i) *The group Γ is geometrically finite.*
(ii) *Every limit point of Γ is either conical or cusped.*
(iii) *Every Dirichlet fundamental polyhedron for Γ is geometrically finite.*

One of Bowditch's definitions of a geometrically finite group is that a group is geometrically finite if and only if all its limit points are either conical or cusped. Thus our definition is equivalent to Bowditch's definitions.

Geometrically finite groups have many finiteness properties. For example, every geometrically finite group is finitely generated. The next theorem [41] describes a new finiteness property of geometrically finite groups.

THEOREM 5.2. *Let Γ be a nonelementary, geometrically finite, discrete group of isometries of B^n such that Γ leaves no m-plane of B^n invariant for $m < n - 1$. Then the group of isometries of B^n/Γ is finite.*

Examples given in [40] show that none of the hypothesis: (1) nonelementary, (2) geometrically finite, nor (3) leaves no m-plane invariant for $m < n - 1$ can be dropped from the theorem. In the case $n = 3$, the last hypothesis is subsumed in the first, and so we have the following corollary.

COROLLARY 5.1. *If Γ is a nonelementary, geometrically finite, discrete group of isometries of B^3, then B^3/Γ has a finite group of isometries.*

We now turn our attention to geometrically finite hyperbolic manifolds.

DEFINITION 5.8. A hyperbolic n-manifold M is *geometrically finite* if M is isometric to the orbit space B^n/Γ of a geometrically finite, torsion-free, discrete group Γ of isometries of B^n.

Let $M = B^n/\Gamma$ be a hyperbolic manifold and let $C(\Gamma)$ be the hyperbolic convex hull of the limit set of Γ. Then $C(\Gamma) \cap B^n$ is a closed, convex, Γ-invariant subset of B^n. The *convex core* of M is the set

$$C(M) = \big(C(\Gamma) \cap B^n\big)/\Gamma.$$

The convex core $C(M)$ is a geodesically connected closed subset of M. If M has finite volume, then $C(M) = M$.

The next theorem was proved by Thurston [48]. See also Bowditch [7] and §12.6 of Ratcliffe [39].

THEOREM 5.3. *A hyperbolic manifold $M = B^n/\Gamma$ is geometrically finite if and only if the r-neighborhood $N(C(M), r)$ of $C(M)$ in M has finite volume for each $r > 0$.*

We deduce immediately the following corollary.

COROLLARY 5.2. *Every hyperbolic manifold of finite volume is geometrically finite.*

We next consider the geometry of the ends of a geometrically finite manifold $M = B^n/\Gamma$. The *ordinary set* $O(\Gamma)$ of Γ is the complement in S^{n-1} of the limit set $L(\Gamma)$. The group Γ acts discontinuously on $O(\Gamma)$ and $O(\Gamma)/\Gamma$ is an $(n-1)$-dimensional manifold called the *ideal boundary* of M. Let $\overline{M}$ denote the union of M and its ideal boundary.

Let Γ be a torsion-free, elementary, discrete group of isometries of U^n of parabolic type and let U be a cusped region in $\overline{U}^n$ for Γ. An incomplete hyperbolic n-manifold, with ideal boundary, equivalent to U/Γ is called an n-dimensional *cusp*.

The following theorem describes the geometry of the ends of a geometrically finite hyperbolic manifold.

THEOREM 5.4. *If M is a geometrically finite hyperbolic n-manifold, then there is a compact n-manifold-with-boundary $\overline{M}_0$ in $\overline{M}$ such that $\overline{M} - \overline{M}_0$ is the disjoint union of a finite number of cusps.*

If M is a hyperbolic manifold of finite volume, then the ideal boundary of M is empty and so we have the following corollary.

COROLLARY 5.3. *If M is a hyperbolic n-manifold of finite volume, then there is a compact n-manifold-with-boundary M_0 in M such that $M - M_0$ is the disjoint union of a finite number of cusps.*

The geometry of geometrically infinite hyperbolic n-manifolds, with finitely generated fundamental group and $n \geqslant 4$, can be very complicated. For example, Kapovich and Potyagailo [25] and Bowditch and Mess [8] have constructed examples of hyperbolic 4-manifolds whose fundamental groups are finitely generated but not finitely presented. It remains a challenge to find the right conditions on a finitely generated, torsion-free, discrete group of isometries of B^n, with $n \geqslant 4$, so that $\overline{M} = (H^n \cup O(\Gamma))/\Gamma$ satisfies a generalization of Ahlfors' finiteness theorem.

References

[1] W. Abikoff, *The uniformization theorem*, Amer. Math. Monthly **88** (1981), 574–592.

[2] C. Adams, M. Hildebrand and J. Weeks, *Hyperbolic invariants of knots and links*, Trans. Amer. Math. Soc. **326** (1991), 1–56.

[3] L. Ahlfors, *Finitely generated Kleinian groups*, Amer. J. Math. **86** (1964), 413–429.

[4] B.N. Apanasov and I.S. Gutsul, *Greatly symmetric totally geodesic surfaces and closed hyperbolic 3-manifolds which share a fundamental polyhedron*, Topology' 90, B. Apanasov, W.D. Neumann, A.W. Reid, and L. Siebenmann, eds, de Gruyter, Berlin (1992), 37–53.

[5] L. Bers, *Finite dimensional Teichmüller spaces and generalizations*, Bull. Amer. Math. Soc. **5** (1981), 131–172.

[6] F. Bonahon, *Bouts des variétés hyperboliques de dimension 3*, Ann. Math. **124** (1986), 71–158.

[7] B.H. Bowditch, *Geometrical finiteness for hyperbolic groups*, J. Funct. Anal. **113** (1993), 245–317.

[8] B.H. Bowditch and G. Mess, *A 4-dimensional Kleinian group*, Trans. Amer. Math. Soc. **344** (1994), 391–405.

[9] P.J. Callahan and A.W. Reid, *Hyperbolic structures on knot complements*, Chaos, Solitons Fractals **9** (1998), 705–738.

[10] R.D. Canary, *Ends of hyperbolic 3-manifolds*, J. Amer. Math. Soc. **6** (1993), 1–35.

[11] J.W. Cannon, *The combinatorial Riemann mapping theorem*, Acta Math. **173** (1994), 155–234.

[12] J.W. Cannon and D. Cooper, *A characterization of cocompact hyperbolic and finite-volume hyperbolic groups in dimension three*, Trans. Amer. Math. Soc. **330** (1992), 419–431.

[13] C. Cao and R. Meyerhoff, *The orientable cusped hyperbolic 3-manifolds of minimum volume*, Invent. Math., to appear.

[14] L.S. Charlap, *Bieberbach Groups and Flat Manifolds*, Universitext, Springer, New York (1986).
[15] S.-S. Chern, *On curvature and characteristic classes of a Riemann manifold*, Abh. Math. Sem. Univ. Hamburg **20** (1955), 117–126.
[16] M.W. Davis, *A hyperbolic* 4*-manifold*, Proc. Amer. Math. Soc. **93** (1985), 325–328.
[17] R. Fricke and F. Klein, *Vorlesungen über die Theorie der automorphen Functionen*, Vols. I, II, B.G. Teubner, Leipzig (1897, 1912).
[18] M. Gromov, *Hyperbolic manifolds, groups and actions*, Riemann Surfaces and Related Topics, I. Kra and B. Maskit, eds, Ann. Math. Studies, Vol. 97 (1980), 183–213.
[19] M. Gromov, *Hyperbolic manifolds according to Thurston and Jørgensen*, Séminaire Bourbaki, Vol. 1979/80, Lecture Notes in Math., Vol. 842, Springer, Berlin (1981), 40–53.
[20] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, Math. Sci. Res. Publ., Vol. 8, S.M. Gersten, ed., Springer, New York (1987), 75–263.
[21] M. Gromov and I. Piatetski-Shapiro, *Non-arithmetic groups in Lobachevsky spaces*, Inst. Hautes Études Sci. Publ. Math. **66** (1988), 93–103.
[22] W. Hantzsche and H. Wendt, *Dreidimensionale euklidische Raumformen*, Math. Ann. **110** (1935), 593–611.
[23] H. Hopf, *Zum Clifford–Kleinschen Raumproblem*, Math. Ann. **95** (1926), 313–339.
[24] H. Hopf, *Die Curvatura integra Clifford–Kleinscher Raumformen*, Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl. (1926), 131–141.
[25] M. Kapovich and L. Potyagailo, *On the absence of Ahlfors' finiteness theorem for Kleinian groups in dimension three*, Topology Appl. **40** (1991), 83–91.
[26] R. Kellerhals, *On the volumes of hyperbolic* 5*-orthoschemes and the Trilogarithm*, Comment. Math. Helv. **67** (1992), 648–663.
[27] R. Kellerhals and T. Zehrt, *The Gauss–Bonnet formula for hyperbolic manifolds of finite volume*, Geometriae Dedicata **84** (2001), 49–62.
[28] W. Killing, *Ueber die Clifford–Klein'schen Raumformen*, Math. Ann. **39** (1891), 257–278.
[29] F. Klein, *Neue Beiträge zur Riemannschen Functionentheorie*, Math. Ann. **21** (1883), 141–218.
[30] F. Klein, *Zur Nicht-Euklidischen Geometrie*, Math. Ann. **37** (1890), 544–572.
[31] J.-S. Li and J.J. Millson, *On the first Betti number of a hyperbolic manifold with an arithmetic fundamental group*, Duke Math. J. **71** (1993), 365–401.
[32] A. Marden, *The geometry of finitely generated Kleinian groups*, Ann. of Math. **99** (1974), 383–462.
[33] S.V. Matveev and A.T. Fomenko, *Constant energy surfaces of Hamiltonian systems, enumeration of three-dimensional manifolds in increasing order of complexity, and computation of volumes of closed hyperbolic manifolds*, Russian Math. Surveys **43** (1988), 3–24.
[34] R. Meyerhoff, *Geometric invariants for* 3*-manifolds*, Math. Intelligencer **14** (1992), 37–53.
[35] J.J. Millson, *On the first Betti number of a constant negatively curved manifold*, Ann. of Math. **104** (1976), 235–247.
[36] Y.N. Minsky, *On rigidity, limit sets, and end invariants of hyperbolic* 3*-manifolds*, J. Amer. Math. Soc. **7** (1994), 539–588.
[37] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Ann. of Math. Studies, Vol. 78, Princeton Univ. Press, Princeton (1973).
[38] G. Prasad, *Strong rigidity of Q-rank* 1 *lattices*, Invent. Math. **21** (1973), 255–286.
[39] J. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Graduate Texts in Math., Vol. 149, Springer, Berlin (1994).
[40] J. Ratcliffe, *On the isometry groups of hyperbolic manifolds*, The Mathematical Legacy of Wilhelm Magnus, Groups, Geometry, and Special Functions, W. Abikoff, J.S. Birman and K. Kuiken, eds, Contemp. Math., Vol. 169 (1994), 491–495.
[41] J. Ratcliffe, *On the isometry groups of hyperbolic orbifolds*, Geometriae Dedicata **78** (1999), 63–67.
[42] J. Ratcliffe and S. Tschantz, *Volumes of integral congruence hyperbolic manifolds*, J. Reine Angew. Math. **488** (1997), 55–78.
[43] J. Ratcliffe and S. Tschantz, *Gravitational instantons of constant curvature*, Class. Quantum Grav. **15** (1998), 2613–2627.
[44] J. Ratcliffe and S. Tschantz, *The volume spectrum of hyperbolic* 4*-manifolds*, Experiment. Math. **9** (2000), 101–125.

[45] B. Riemann, *Über die Hypothesen, welche der Geometrie zu Grunde liegen*, Abh. Ges. Wiss. Göttingen **13** (1867), 133–152.

[46] C.L. Siegel, *Über die analytische Theorie der quadratischen Formen* II, Ann. Math. **37** (1936), 230–263.

[47] P. Scott, *The geometries of* 3*-manifolds*, Bull. London Math. Soc. **15** (1984), 401–487.

[48] W.P. Thurston, *The Geometry and Topology of Three-Manifolds*, Lecture Notes, Princeton Univ. (1979).

[49] W.P. Thurston, *Three dimensional manifolds, Kleinian groups, and hyperbolic geometry*, Bull. Amer. Math. Soc. **6** (1982), 357–381.

[50] W.P. Thurston, *Hyperbolic structures on* 3*-manifolds* I: *Deformation of acylindrical manifolds*, Ann. Math. **124** (1986), 203–246.

[51] W.P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. **19** (1988), 417–431.

[52] W.P. Thurston, *Three-Dimensional Geometry and Topology, Vol. 1*, S. Levy, ed., Princeton Univ. Press, Princeton (1997).

[53] A. van der Poorten, *A proof that Euler missed ... Apéry's proof of the irrationality of* $\zeta(3)$, Math. Intelligencer **1** (1979), 195–203.

[54] E.B. Vinberg and O.V. Shvartsman, *Discrete groups of motions of spaces of constant curvature*, Geometry II, E.B. Vinberg, ed., Encyclopaedia Math. Sci., Vol. 29, Springer, Berlin (1993), 139–254.

[55] H.-C. Wang, *Topics on totally discontinuous groups*, Symmetric Spaces, W.M. Boothby and G.L. Weiss, eds, Pure Appl. Math., Vol. 8, Marcel Dekker, New York (1972), 459–487.

[56] J.R. Weeks, *Convex hulls and isometries of cusped hyperbolic* 3*-manifolds*, Topology Appl. **52** (1993), 127–149.

[57] N. Wielenberg, *Hyperbolic* 3*-manifolds which share a fundamental polyhedron*, Riemann Surfaces and Related Topics, I. Kra and B. Maskit, eds, Ann. of Math. Studies, Vol. 97, Princeton Univ. Press, Princeton (1980), 505–513.

[58] J.A. Wolf, *Spaces of Constant Curvature*, 5th edn, Publish or Perish, Wilmington (1984).

CHAPTER 18

Heegaard Splittings of Compact 3-Manifolds

Martin Scharlemann

Department of Mathematics, University of California, Santa Barbara, CA 93106, USA

Contents

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Background

Here is a simple way to build a complicated 3-manifold. Begin with the 3-ball B^3 and in its boundary pick out two disjoint 2-disks D_0 and D_1. Using those disks, attach to B^3 a *handle*, that is a copy of $D^2 \times I$, by identifying $D^2 \times \{i\}$ with D_i, $i = 0, 1$. Depending on the orientation with which the ends of the handle are attached, the result is either $D^2 \times S^1$ or the non-orientable disk bundle over S^1, bounded by the Klein bottle. One can continue to add more handles to B^3 in a similar way. The result of adding g of them is called a *genus g handlebody*. Topologically, there are exactly two of them for any g, one of them orientable and the other not orientable, for once a non-orientable handle is attached, the end of any other handle can be slid over it, converting an orientable handle to a non-orientable, and vice versa. These manifolds are easily understood and not yet very complicated.

Now imagine taking two such handlebodies, H_1 and H_2, of the same genus and orientability. Then ∂H_1 and ∂H_2 are homeomorphic (the connected sum of g tori or Klein bottles) and one can construct a complicated 3-manifold by attaching H_1 to H_2 by a possibly complicated homeomorphism of their boundaries. The resulting closed 3-manifold M can be written $M = H_1 \cup_S H_2$, where S is the surface ∂H_i in M. This structure on M is called a *Heegaard splitting* of M and S is a *splitting surface* (of a Heegaard splitting). Two Heegaard splittings of M are *isotopic* if their splitting surfaces are isotopic in M. They are *homeomorphic* if there is a homeomorphism of M carrying the splitting surface of one to the splitting surface of the other.

This method of constructing 3-manifolds is attributed to Heegaard [18] (see [35] for a translation into English of the relevant parts) though it was probably known to Poincare.

Natural questions arise: How *universal* is this construction? That is, how many closed 3-manifolds have such a structure? Is there a natural extension to 3-manifolds with boundary? This question is considered in Section 2. How *unique* is such a structure? That is, given two such structures on the same 3-manifold, how are they related? This question is addressed in Sections 6 and 7. How *useful* is the structure? That is, what information about the 3-manifold can be gleaned from the structure of a Heegaard splitting. Such questions are addressed in 5 and 8.

A useful earlier survey of the subject is [53], which focuses on Heegaard diagrams and on group presentations (briefly discussed in Sections 2.3 and 5 below). I've relied heavily on its historical account. A central recent development has been an understanding of the importance of *strongly irreducible* Heegaard splittings (see 3.3), so their role has been chosen as a major theme of this survey.

2. Heegaard splittings and their guises

2.1. *Splittings from triangulations*

A foundational theorem of Moise [29] (see also [3]) says that all 3-manifolds can be triangulated. That is, given a compact connected 3-manifold M there is a finite simplicial complex K which is homeomorphic to M. For our purposes there are two important connected finite graphs in such a triangulation K: the 1-skeleton K^1 and the *dual* 1-skeleton

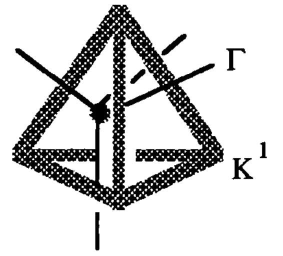


Fig. 1.

Γ, defined as follows. (See Figure 1.) The vertices of Γ are the barycenters of the 2- and 3-simplices of K and edges connect the barycenter of a 3-simplex to the barycenter of each of its faces.

In case M is closed, each 2-simplex is the face of precisely two 3-simplices, so each vertex in Γ coming from a 2-simplex has valence 2. The edges of Γ incident to such a vertex can therefore be amalgamated into a single edge, so that Γ becomes a graph in which each vertex corresponds to a 3-simplex and each edge to the 2-simplex which it intersects. Now the regular neighborhood of a finite graph in a 3-manifold is easily seen to be a handlebody of genus $|edges| - |vertices| + 1$, for a regular neighborhood of a maximal tree is just a 3-ball, and a regular neighborhood of each remaining edge contributes a 1-handle. Furthermore, the region between regular neighborhoods of K^1 and Γ is a product region, as can be seen easily in each 3-simplex. So, after thickening the regular neighborhoods of these graphs, M can be viewed as the union of a regular neighborhood of K^1 and a regular neighborhood of Γ along their common boundary. This is a Heegaard splitting of M.

2.2. *Splitting 3-manifolds with boundary*

The construction of Heegaard splittings for closed 3-manifolds in Section 2.1 suggests several possible ways of extending the definition of Heegaard splitting to cover the case in which the 3-manifold has boundary. The most useful is the following: Write ∂M as the disjoint union of two sets of components, $\partial_1 M$ and $\partial_2 M$. Choose a triangulation K fine enough so that no simplex is incident to more than one boundary component. Let K' be its barycentric subdivision. Delete the interior of all simplices of K' incident to $\partial_2 M$. The resulting 3-manifold M' is homeomorphic to M, since only a collar of $\partial_2 M$ has been removed; let $\partial_2' M$ denote $\partial_2 M$ in this new triangulation. (Then $\partial_2' M$ contains the subcomplex of Γ determined by simplices incident to $\partial_2 M$.) Let $\Gamma_1 \subset M'$ be the union of $\partial_1 M$ and all vertices and edges of K not incident to $\partial_2 M$. Let Γ_2 be the union of $\partial_2' M$ and all vertices and edges of the dual 1-complex $\Gamma \cap M'$ not incident to Γ_1. Again it's easy to check that M is the union of a regular neighborhood of the complexes Γ_1 and Γ_2 along their homeomorphic boundary, which is still a closed connected surface.

This construction suggests the following way of defining a Heegaard splitting on a 3-manifold with boundary. A *compression body* H is a connected 3-manifold obtained from

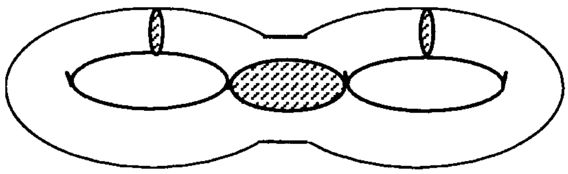


Fig. 2.

a closed surface $\partial_- H$ by attaching 1-handles to $\partial_- H \times \{1\} \subset \partial_- H \times I$. (It is conventional to consider a handlebody to be a compression body in which $\partial_- H = \emptyset$.) Dually, a compression body is obtained from a connected surface $\partial_+ H$ by attaching 2-handles to $\partial_+ H \times \{1\} \subset \partial_+ H \times I$ and 3-handles to any 2-spheres thereby created. The cores of the 2-handles are called *meridian disks* and a collection of meridian disks is called *complete* if each of its complementary components is either a ball or $\partial_- H \times I$ (see Figure 2 for H a handlebody). Suppose two compression bodies H_1 and H_2 have $\partial_+ H_1 \simeq \partial_+ H_2$. Then glue H_1 and H_2 together along $\partial_+ H_i = S$. The resulting compact 3-manifold M can be written $M = H_1 \cup_S H_2$ and this structure also is called a *Heegaard splitting* of the 3-manifold with boundary M (or, more specifically, of the triple $(M; \partial_- H_1, \partial_- H_2)$). It follows from the motivating discussion above that every compact 3-manifold has a Heegaard splitting.

2.3. *Splittings as handle decompositions – Heegaard diagrams*

Those familiar with handle decompositions of compact n-manifolds (see, e.g., [36, Chapter 6] for the notation and viewpoint used here) will recognize similarities between the ways in which Heegaard splittings and handle decompositions are derived from a triangulation. The similarity goes deeper. Suppose $H_1 \cup_S H_2$ is a Heegaard splitting of a 3-manifold $(M; \partial_1 M, \partial_2 M)$. Then H_1 is obtained from $\partial_1 M \times I$ by attaching 1-handles and H_2 is obtained from $S = \partial_+ H_1 = \partial_+ H_2$ by attaching 2- and 3-handles. From this point of view a Heegaard splitting is just a standard handle decomposition of M viewed as a cobordism between $\partial_1 M$ and $\partial_2 M$.

There is an advantage to this point of view. It is a standard trick in handle theory that the order of handles can frequently be rearranged. Always r-handles can be attached before $(r + 1)$-handles, so that handles can be attached in ascending order. It is not generally true that an $(r + 1)$-handle can be attached before an r-handle – it's necessary and sufficient that the attaching r-sphere of the $(r + 1)$-handle be disjoint from the belt $(n - r - 1)$-sphere of the r-handle. Translated into the language of Heegaard splittings this means that the natural order of handles can be rearranged if and only if there are essential disks in H_1 and H_2 whose boundaries are disjoint in S. This is a situation whose importance we will discuss later (see 3.3).

In this handle picture, all the topological information is contained in an understanding of the 1- and 2-handles, since the remaining 3-handles (if any) are uniquely determined by the spherical components of the boundary. Encouraged by this observation, we look for an efficient way of describing the way in which 2-handles are attached. We consider the case in which M is closed; if M has boundary the situation is analogous but a bit more complicated. When M is closed, H_1 is a genus g handlebody. The attaching curves $\partial \Delta_2$ for the cores Δ_2 of the 2-handles constitute a family of simple closed curves in ∂H_1. We may as well isotope $\partial \Delta_2$ to intersect a chosen minimal complete collection Δ_1 of meridian disks for H_1 transversally and minimally. When H_1 is cut open along Δ_1 it becomes a 3-ball B^3, on whose boundary appear two copies of each disk of Δ_1. Let $V \subset \partial B^3$ be this collection of disks. The attaching curves $\partial \Delta_2$ are (typically) also cut up – into a collection $\mathcal{A}$ of arcs and simple closed curves in $\partial B^3 - V$. The ends of each arc in $\mathcal{A}$ lie in ∂V.

If the splitting is irreducible (see (see 3.2), note that $\mathcal{A}$ consists entirely of arcs, since any simple closed curve in $\mathcal{A}$ bounds a disk in B^3 and the union of this disk and a 2-handle core with the same boundary would give a reducing sphere. When no component of $\mathcal{A}$ is a simple closed curve, we can think of the union of V and $\mathcal{A}$ as defining a graph Γ in ∂B^3, with fat vertices V and edges $\mathcal{A}$. There is additional structure, of course, which identifies each pair of vertices of V that began as the same disk in Δ_1 and also identifies ends of edges that were cut at $\partial \Delta_1$. The graph Γ has some pleasant properties. For example, there are no trivial loops in Γ, for such a loop could have been removed by an isotopy of $\partial \Delta_2$ that lowers $\partial \Delta_1 \cap \partial \Delta_2$. But there are a lot of choices made in the construction or Γ (e.g., Δ_1 and Δ_2) so it is not particularly well-defined. The use of these diagrams to study the underlying 3-manifold can be quite complicated and is often disappointing.

Sometimes the ability to rechoose Δ_1 and Δ_2 can be useful. For example, although we have observed that Γ contains no trivial loops, it is also true, when the splitting is irreducible, that if any loop at all appears, the diagram can be simplified. A loop in a Heegaard diagram (i.e., an edge in $\mathcal{A}$ both of whose ends lie on the same vertex in V) is sometimes called a *wave*. A wave, and the vertex v in V on which it is based, divides ∂B^3 into two disks. One of them (call it E) does not contain the other vertex in V that is identified with v in Δ_1. All ends of 1-handles in H_1 represented by vertices that lie in E can be dragged over the 1-handle in H_1 whose cocore is v (thereby redefining Δ_1). At this point the wave becomes an inessential loop, which can be removed by an isotopy. The net effect is to reduce $\Delta_1 \cap \Delta_2$ by redefining Δ_1.

We refer the reader to the excellent [53] for a more thorough discussion of Heegaard diagrams.

2.4. *Splittings as Morse functions and as sweep-outs*

Smooth manifolds admit handle structures just as PL manifolds do. One way of showing this classical fact is via Morse theory [28]. A generic smooth height function h from the smooth manifold M to R will have only non-degenerate critical points. At each critical height t_0, as the part $h^{-1}(-\infty, t_0 - \varepsilon]$ of M lying below t_0 changes to $h^{-1}(-\infty, t_0 + \varepsilon]$, the topological effect is to add a handle. The handles can then be rearranged to appear in ascending order, just as in the PL theory.

The argument is reversible. Given a handle structure on a smooth manifold M one can easily construct a Morse function which induces that handle structure. So associated to a Heegaard splitting there is also a Morse function. That is, given a Heegaard structure $H_1 \cup_S H_2$ on $(M; \partial_1 M, \partial_2 M)$ there is a Morse function $h : M \to [0, 1]$ with singular values $0 < a_1 < a_2 < \cdots < a_k < b_1 < \cdots < b_j < 1$ so that

(i) a_i is an index one critical level,
(ii) b_j is an index two critical level,
(iii) $h^{-1}(0) = \partial_1 M$ or is an index zero critical point if $\partial_1 M = \emptyset$,
(iv) $h^{-1}(1) = \partial_2 M$ or is an index three critical point if $\partial_2 M = \emptyset$,
(v) if $a_k < t < b_1$ then $h^{-1}(t) \cong S$.

A reason for taking this viewpoint is that it is sometimes advantageous to put knots and graphs in M into the simplest possible position with respect to the Heegaard splitting. One way to accomplish this is to incorporate Gabai's powerful notion of "thin position" into the theory and make the knot or graph thin with respect to this Morse function (cf. [7,41], 3.7, 4.1).

A similar way to use a Heegaard splitting to parameterize the 3-manifold M is to focus attention on heights between the top index-one critical level a_k and the bottom index-two critical level b_1, as we now explain.

Define the *spine* Σ of a handlebody H to be a finite graph in H for which H is a regular neighborhood. From the construction, every handlebody has a spine (for a closed triangulated manifold, the 1-skeleton and the dual 1-skeletons are spines of the relevant handlebodies in the construction described in Section 2.1 above). The spine of H is not uniquely defined, but any two spines differ by a sequence of "edge-slides" (see [41, 1.2]). For H a compression body, a spine Σ is a graph in H so that $\Sigma \cap \partial H = \Sigma \cap \partial_- H$ consists only of valence one vertices and H deformation retracts to $\Sigma \cup \partial_- H$. (See Figure 3.) Again the construction of H guarantees the existence of a spine, and two spines of the same compression body differ by a series of edge slides, where ends of edges may be slid along paths in $\partial_- H$.

Notice that the complement of a spine in H is homeomorphic to $\partial_+ H \times I$. Suppose then we are given a Heegaard splitting $H_1 \cup_S H_2$ of M and spines Σ_i of each H_i. Then $M - (\Sigma_1 \cup \Sigma_2)$ is just a product $S \times I$. This parameterization of $M - (\Sigma_1 \cup \Sigma_2)$ is sometimes called a "sweep-out" by S since S sweeps between one spine and the other. This viewpoint allows great flexibility in the positioning of the splitting surface. (See Section 7.4.)

A related idea is to consider a single spine, say $\Sigma_1 \subset M$, as a graph in M for which there are ∂-singular compressing disks (the meridian disks of Σ_2). In some situations the ∂-singular disks can be used to slide Σ_1 into useful positions. (See [33] and [42].)

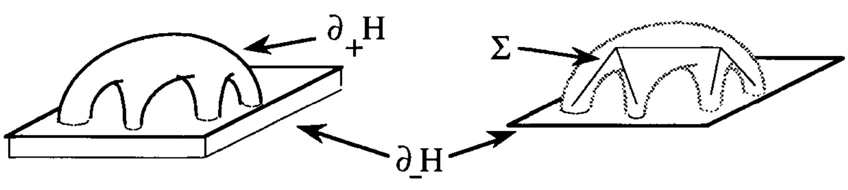


Fig. 3.

3. Structures on Heegaard splittings

3.1. *Stabilization*

As we have seen, Heegaard splittings have connections to triangulations, handle decompositions, and Morse functions on 3-manifolds. Just as triangulations can be subdivided, or a Morse function locally perturbed to introduce cancelling critical points, or, in a handlebody description, a cancelling pair of handles can be inserted, so there is a natural and trivial way of making a Heegaard splitting more complicated. Suppose $H_1 \cup_S H_2$ is a Heegaard splitting of a 3-manifold M and α is a properly imbedded arc in H_2 parallel to an arc in S. Here "parallel" means that there is an embedded disk D in H_2 whose boundary is the union of α and an arc in $\partial_+ H_2$. Now add a neighborhood of α to H_1 and delete it from H_2. This adds a 1-handle to H_1 (whose core is α) and, topologically, also adds a 1-handle to H_2 (whose cocore is D). So once again the result is a Heegaard splitting $H_1' \cup_{S'} H_2'$, where the genus of each H_i' is one greater than H_i. This process is called a *stabilization* of S.

Stabilization is uniquely defined. That is, any two splittings obtained by stabilizing the same splitting surface are isotopic. On the other hand, there is no reason to believe that if the stabilizations of two different splitting surfaces are isotopic, then the original surfaces were isotopic. So in stabilizing a splitting surface we may lose and can't gain information about its structure. Interest therefore focuses on splittings which are not stabilizations of other splittings, that is splittings which cannot be destabilized. How is this detectable? (See Figure 4.)

LEMMA 3.1. *A splitting $M = H_1 \cup_S H_2$ can be destabilized if and only if there are properly imbedded disks $D_i \subset H_i$ so that $|\partial D_1 \cap \partial D_2| = 1$.*

PROOF. Suppose a splitting is stabilized as above. Then let D_1 be the cocore disk of the 1-handle attached along α and let D_2 be the disk D.

Conversely, suppose disks $D_i \subset H_i$ are as in the lemma. Because the boundaries intersect in a single point, each disk is non-separating and hence essential. Let T_i be the surface obtained from S by compressing along D_i, converting H_i into a simpler compression body J_i. The union of a bicollar of D_1 in H_1 and D_2 in H_2 along the square in which they intersect is a 3-ball intersecting each of T_i in a hemisphere, and so defines an isotopy between the T_i. In particular T_1 divides M into J_1 and an isotope of J_2 and so is a Heegaard splitting

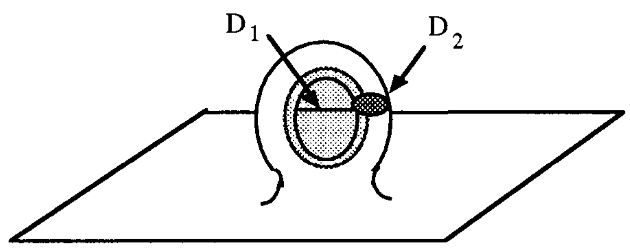


Fig. 4.

surface. It's easy to see that stabilizing T_1 gives S: the 1-handle dual to D_1 corresponds to the arc α and the disk D_2 corresponds to the disk D. □

3.2. *Reducible splittings*

Suppose M and M' are two 3-manifolds with Heegaard splittings $H_1 \cup_S H_2$ and $H_1' \cup_{S'} H_2'$. From these we can naturally construct a Heegaard splitting of the connected sum $M'' = M\#M'$ as follows: Remove from M and M' 3-balls B and B' which intersect S and S' respectively in equatorial disks. Glue together the boundaries of the 3-balls so that each hemisphere $H_i \cap \partial B$ is attached to $H_i' \cap \partial B'$. The resulting surface $S\#S'$ splits $M\#M'$ into compression bodies H_i''. To see that the complementary pieces are compression bodies, note that topologically H_i'' is obtained from the disjoint union of H_i and H_i' by attaching a 1-handle whose two ends lie in $\partial_+ H_i$ and $\partial_+ H_i'$ respectively.

Conversely, given a Heegaard splitting $H_1 \cup_S H_2$ of a 3-manifold M'' and a 2-sphere P which intersects M in a single circle, we can get a Heegaard splitting of the reduced 3-manifold, obtained by doing surgery on P (the reduced manifold is the disjoint union of M and M' when P is separating, as in the above example). If P bounds a ball in M'' and S intersects the ball in a single equatorial disk, then the manifolds M'' and M, say, are the same and get the same splitting. Otherwise, the splitting of the reduced manifold is simpler, since the genus of the splitting surface is reduced, and the Heegaard splitting of M'' can be easily reconstructed from the splitting of the reduced manifold. These considerations lead to the following:

DEFINITION 3.2. A Heegaard splitting $H_1 \cup_S H_2$ is *reducible* if there is a 2-sphere which intersects S in a single essential circle.

An alternate way of saying this is that there are essential properly imbedded disks $D_i \subset H_i$ so that $\partial D_1 = \partial D_2$ in S. (See Figure 5.)

There is a connection with stabilization, given by the following lemma:

PROPOSITION 3.3. *Suppose $H_1 \cup_S H_2$ is a splitting that can be destabilized. Then either it is reducible or it is the standard genus one splitting of S^3.*

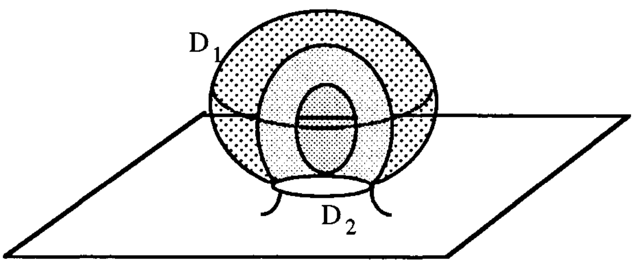


Fig. 5.

PROOF. Let $D_i \subset H_i$ be disks so that $\partial D_1 \cap \partial D_2$ is a single point. As in the proof of 3.1, let B be the union of a bicollar of D_1 in H_1 and a bicollar of D_2 in H_2 along the square in which they intersect. B is a 3-ball whose boundary sphere P can be moved slightly (e.g., increase the radius of both disks slightly) so that P intersects each H_i in a single hemisphere and so that the curve $c = P \cap S$ (the boundary of a regular neighborhood of the figure-eight $\partial D_1 \cup \partial D_2$) cuts off from S a punctured torus. Unless this curve is inessential in S the boundary of the 3-ball is a reducing sphere. If the curve is inessential, then S is a torus dividing M into two solid tori, whose meridians intersect in a single point. This is the genus one Heegaard splitting of S^3. □

One of the first major theorems on Heegaard splittings, due to Haken, is that any Heegaard splitting of a reducible manifold is a reducible splitting. The theorem is important not just for what it says, but for the type of argument which is used.

THEOREM 3.4 ([17]). *Suppose M is a reducible manifold with a Heegaard splitting $H_1 \cup_S H_2$. Then there is a reducing sphere P for M so that $P \cap S$ is a single circle.*

PROOF. There are two ways to do this. Similar to Haken's original proof is that given in [21, II.7]. One can assume that P intersects (either) one of the compression bodies only in disks. (One way to do this is to put a spine of H_2, say, transverse to P and take H_2 to be very thin.) The idea will be to minimize the number of circles of intersection, under the assumption that P intersects one of the compression bodies only in disks. If P intersects H_2 only in disks, consider the planar surface $P_1 = P \cap H_1$. Compress and ∂-compress P_1 as much as possible. Compressions of P_1 will convert P into two spheres, at least one of which is a reducing sphere – restrict attention to that one. At the end of this process P_1 will be converted to a surface P' which is disjoint from a complete collection of meridian disks for H_1 (otherwise curves of intersection can be used to compress or ∂-compress) and, for any essential curve α in $\partial_- H_1$, disjoint from a spanning annulus $\alpha \times I$ (same argument). It follows that $P' \cap H_1$ is a collection of disks. What is not obvious, but can be explicitly calculated, is that the number of disks in $P' \cap H_1$ is lower than the original $P \cap H_2$. The process is continued, switching the roles of H_1 and H_2 until there is only one intersection curve.

Another approach is given in [42]. Put Σ_2, the spine of H_2, transverse to P. Let Δ be a complete collection of compressing disks for H_1 viewed as a ∂-singular collection of disks in the complement of Σ_2. Put Δ transverse to P. Circles of intersection can be removed, just as in the previous argument, so that $(\Sigma_2 \cup \Delta) \cap P$ becomes a graph $\Gamma \subset P$ with vertices $\Sigma_2 \cap P$ and edges $\Delta \cap P$. Trivial loops of Γ can be eliminated at the cost of merely changing Δ, and a vertex incident to some edges but no loops can be used to slide edges of Σ_2 in a way that lowers $\Sigma_2 \cap P$. (This is the hard part to see.) The upshot is that, eventually, there is guaranteed to be an isolated vertex. This picks out a meridian μ of H_1 which is disjoint from a complete collection of meridian disks for H_2. If H_2 is a handlebody this implies that $\partial\mu$ also bounds a meridian in H_2 and so $H_1 \cup_S H_2$ is reducible. If H_2 is merely a compression body, we can only conclude that there is a ∂-reducing disk for M which intersects S in a single curve. But we can surger M along this disk to get a new reducible 3-manifold and continue the process until an appropriate sphere is found. □

The last step of the second proof suggests a new notion:

DEFINITION 3.5. A Heegaard splitting $M = H_1 \cup_S H_2$ is *∂-reducible* if there is a ∂-reducing disk for M which intersects S in a single curve.

It also suggests the following analogue to Theorem 3.4.

PROPOSITION 3.6. *Any Heegaard splitting of a ∂-reducible 3-manifold is ∂-reducible.*

PROOF. Both proofs above easily generalize. □

A more difficult theorem, discussed in more detail in Section 6.1 but relevant here, characterizes Heegaard splittings of the 3-sphere.

THEOREM 3.7 ([52]). *Every positive genus Heegaard splitting of S^3 is stabilized.*

This implies, more fully, that any positive genus Heegaard splitting of S^3 is obtained by stabilizing the unique genus zero splitting into 3-balls. So a Heegaard splitting of S^3 is completely determined by its genus.

Armed with Theorem 3.7 we can prove a sort of converse to 3.3.

THEOREM 3.8. *Suppose M is an irreducible 3-manifold and $H_1 \cup_S H_2$ is a reducible Heegaard splitting of M. Then $H_1 \cup_S H_2$ is stabilized.*

PROOF. Let P be a sphere which intersects S in a single essential circle. Since M is irreducible, P bounds a 3-ball in M, so the manifold obtained by reducing M along P is the disjoint union of S^3 and a homeomorph of M. The induced Heegaard splitting of the former is, by 3.7, stabilized. Its stabilizing disks, when viewed back in $H_1 \cup_S H_2$ show that S was also stabilized. □

3.3. *Weakly reducible splittings*

In 1987 [12] Casson and Gordon discovered a new structure on Heegaard splittings which is perhaps less natural than those described above but which has turned out to be quite useful.

DEFINITION 3.9. A Heegaard splitting $H_1 \cup_S H_2$ is *weakly reducible* if there are essential disks $D_i \subset H_i$ so that ∂D_1 and ∂D_2 are disjoint in S.

REMARKS.

(i) This notion coincides precisely to the assertion that, in viewing the Heegaard structure as a handle decomposition, at least one 2-handle (D_2) can be attached before all 1-handles are attached (in particular the 1-handle dual to D_1).

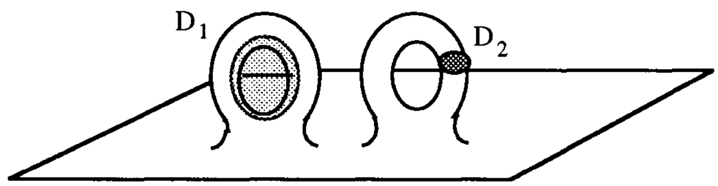


Fig. 6.

(ii) Any reducible Heegaard splitting is weakly reducible, simply by cutting the sphere P that intersects S into two disks along $P \cap S$, then pushing the two boundaries apart.

(iii) A splitting that is not weakly reducible is called *strongly irreducible*.

Here are two sample applications of this structure:

LEMMA 3.10 ([42]). *Suppose $H_1 \cup_S H_2$ is a strongly irreducible splitting of a 3-manifold M and F is a disk in M transverse to S with $\partial F \subset S$. Then ∂F also bounds a disk in some H_i.*

PROOF. The proof is by induction on $|S \cap int(F)|$. If the interior of F is disjoint from S there is nothing to prove. If $S - F$ has any disk components D then, by replacing the subdisk of F bounded by ∂D by a parallel copy of D we can decrease $|S \cap int(F)|$. So assume that each curve in $S \cap F$ is essential in S.

A disk component of $F - S$ compresses S in one of the two compression bodies, say H_1. Then by strong irreducibility of S, all disk components of $F - S$ lie in H_1. If any pair of curves of $F \cap S$ are nested and inessential in F then the outer curve of the innermost such pair cuts off a component P of $F - S$ so that all but one of the curves in ∂P are adjacent to disks in H_1 (hence $P \subset H_2$) and precisely one, denoted α, is not. Compress S into H_1 along 2-handles whose cores are the disks with boundaries on ∂P. Let M_- be the 3-manifold obtained from H_2 by attaching these 2-handles to H_2. Then $\alpha \subset \partial M_-$ is inessential in M_- so, by strong irreducibility and 3.6, α is inessential in ∂M_-. Push the disk α bounds in ∂M_- slightly into H_1 and observe that this is then a disk D in H_1 whose boundary is parallel to α in the component of F adjacent to P across α. Replacing the subdisk of F bounded by α (or all of F if $\alpha = \partial F$) with D lowers $|S \cap int(F)|$. □

THEOREM 3.11 ([12]). *If $M = H_1 \cup_S H_2$ is a weakly reducible splitting then either $H_1 \cup_S H_2$ is reducible or M contains an incompressible surface.*

PROOF. S can be compressed simultaneously in both directions, that is, both into H_1 and simultaneously into H_2. Let $\Delta_1 \subset H_1$ and $\Delta_2 \subset H_2$ be collections of disjoint meridians in the respective compression bodies so that $\partial \Delta_1$ and $\partial \Delta_2$ are disjoint in S and the families Δ_i are maximal with respect to this property. That is, if S_i represents the surface in H_i obtained by compressing S along Δ_i, then any further compressing disks of S_i into H_i will necessarily have boundaries intersecting the boundaries of the other disk family (or any obtained from it by 2-handle slides – a requirement that makes the definition of "maximal" here mildly subtle).

Let $\bar{S}$ be the surface obtained by compressing S_1 along Δ_2 (or, symmetrically, S_2 along Δ_1). $\bar{S}$ separates M into the remnant W_1 of H_1 and the remnant W_2 of H_2. Dually, H_1 can be recovered from W_1 by removing some "tunnels" (neighborhoods of arcs) from W_1 and attaching some 1-handles in W_2. A helpful and vivid picture is to imagine H_1 red and H_2 blue. The compressions of S to $\bar{S}$ along the Δ_i cover $\bar{S}$ with both red and blue spots, two red spots for each disk in Δ_1 and two blue spots for each disk in Δ_2. S is recovered from $\bar{S}$ by attaching red tubes in W_2 with ends on red spots and blue tubes in W_1 with ends on blue spots.

The surface $\bar{S}$ is incompressible in M. To see this, suppose that $\bar{S}$ compresses into W_1, say. After pushing $\bar{S}$ slightly into W_2, we can view S_1 as a Heegaard splitting surface of W_1, that is $W_1 = H_1 \cup_{S_1} (W_1 \cap H_2)$. The compression of $\bar{S}$ is a ∂-reduction of W_1. By Theorem 3.6 there is a ∂-reducing disk D that intersects S_1 in a single circle. We can take ∂D to be disjoint from the "red spots" (i.e., disjoint from Δ_1) and, after some 2-handle slides among the Δ_2, we can make Δ_2 disjoint from the annulus $D - H_1$. But then $D \cap H_1$ makes S_1 compressible in H_1 via a disk disjoint from Δ_2, contradicting the maximality of Δ_1.

Unless $\bar{S}$ is a collection of spheres, we are through. Suppose $\bar{S}$ is a collection of spheres. Note that at least one, $\bar{S}_0$, has both a red spot and a blue spot. For otherwise, when S is recovered from $\bar{S}$ by attaching red and blue tubes, S would consist of two components: one containing all red tubes and one containing all blue. Choose in $\bar{S}_0$ a simple closed curve that separates in the sphere $\bar{S}_0$ the red spots from the blue spots. Push the interior of the disk in $\bar{S}_0$ that contains the red spots (respectively blue spots) completely into H_1 (respectively H_2). Then $\bar{S}_0$ is the union of a red disk and a blue disk along a curve, i.e., it is a reducing sphere for the original Heegaard splitting. □

Note that at the end of the proof above we have $\bar{S}$ dividing M into two (not necessarily connected) 3-manifolds, W_1 and W_2. Each component of W_i inherits a Heegaard splitting surface (a component of S_i) of lower genus than S. This splitting itself may be weakly reducible and we can continue the process. Ultimately an irreducible Heegaard splitting $M = H_1 \cup_S H_2$ is thereby broken up into a series of strongly irreducible splittings (see [43]). That is, we can begin with the handle structure determined by $H_1 \cup_S H_2$ and rearrange the order of the 1- and 2-handles, so that ultimately

$$M = M_0 \cup_{\bar{S}_1} M_1 \cup_{\bar{S}_2} \cdots \cup_{\bar{S}_m} M_m.$$

The 1- and 2-handles which occur in M_i provide it with a strongly irreducible splitting (in each component) $A_i \cup_{P_i} B_i$ with $\partial_- A_i = \bar{S}_i$, $\partial_- B_{i-1} = \bar{S}_i$ for $1 \leqslant i \leqslant m$, $\partial_- A_0 = \partial_- H_1 \subset \partial M$, $\partial_- B_m = \partial_- H_2 \subset \partial M$. Each component of each $\bar{S}_i$ is a closed incompressible surface of positive genus and, for any i, only one component of M_i is not a product. None of the compression bodies $A_i, B_{i-1}, 1 \leqslant i \leqslant m$, is trivial. If $\partial_- A$ or $\partial_- B$ is compressible in M (so in particular M is ∂-reducible) then respectively A_0 or B_m is trivial (i.e., just a product). Such a rearrangement of handles will be called an *untelescoping* of the Heegaard splitting.

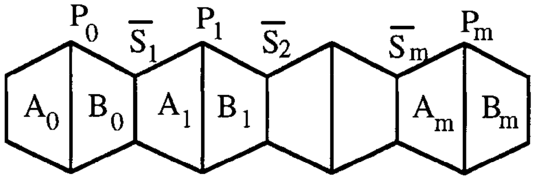


Fig. 7.

So we see that just as a reducible splitting can be broken up by spheres into a connected sum of irreducible Heegaard splittings so irreducible Heegaard splittings can, by rearranging handles, be decomposed by incompressible surfaces into a sequence of strongly irreducible splittings. In effect, strongly irreducible splittings can be viewed as the fundamental building blocks of general Heegaard splittings.

The inverse process is also of interest. Suppose an incompressible surface $\bar{S}$ divides a connected 3-manifold M into two pieces M_0 and M_1 and, for $i = 0, 1$ there are surfaces $P_i \subset M_i$ which divide (each component of) M_i into compression bodies A_i and B_i, with $\partial_- B_0 = \bar{S} = \partial_- A_1$. From this we can recover a Heegaard splitting of M by a process called *amalgamation* (see [44]). Informally, we regard the two Heegaard splittings as handle decompositions and rearrange the handles so that all the 1-handles are attached to $\partial_- A_0$ before the 2-handles are attached. More formally, do the following: The 3-manifold $B_0 \cup_{\bar{S}} A_1$ can be viewed as obtained from $\bar{S} \times [-1, 1]$ by attaching some 1-handles (from B_0) to $\bar{S} \times \{-1\}$ and some 1-handles (from A_1) to $\bar{S} \times \{1\}$. The attaching disks of these 1-handles, in $\bar{S} \times \{\pm 1\}$ can be taken to project to disjoint disks in $\bar{S}$. Collapse $\bar{S} \times I$ to $\bar{S}$. Then the 1-handles of B_0 are attached to $P_1 = \partial_+ B_1$, which makes it a compression body B, and the 1-handles of A_1 are attached to $P_0 = \partial_+ A_0$, which makes it a compression body A. Moreover, $\partial_+ A = \partial_+ B$. If we denote this surface S, then $M = A \cup_S B$ is a Heegaard splitting.

4. Heegaard splittings in nature: Seifert manifolds

An important class of irreducible 3-manifolds is the Seifert manifolds. We restrict our comments here to those Seifert manifolds constructed only with orientation preserving data, as it is these on which Heegaard splittings are best understood. We call these *fully orientable* Seifert manifolds.

Here is how a fully orientable Seifert manifold is constructed (see also [47]). Begin with F a compact orientable surface (called the base surface of the Seifert manifold) and choose n points $x_1, \ldots, x_n \in F$. Choose small disjoint disks E_i around the x_i. Let $F_- = F - \bigcup(\bigcup_i \partial E_i)$. In $F \times S^1$ do Dehn surgery on each $E_i \times S^1$ as follows (see [11]): For each $1 \leqslant i \leqslant n$ remove $E_i \times S^1$ and glue back a solid torus T_i so that a circle $\{x\} \times S^1 \subset \partial E_i \times S^1 \subset \partial F_i \times S^1$ is identified with a (p_i, q_i) torus knot on ∂T_i. This means a knot going $p_i \geqslant 2$ times around the longitude (i.e., crossing a meridian p_i times) and q_i times around a meridian (i.e., crossing a longitude q_i times). Because of ambiguity in the choice of longitude for T_i, q_i is only defined mod p_i and it is customary to take $1 \leqslant q_i < p_i$. Once this Dehn surgery is done, the projection of $\partial E_i \times S^1$ to B^2 extends to a projection of T_i to E_i in which the inverse image of each point of $E_i - \{x_i\}$ is a (p_i, q_i) torus knot in T_i and the inverse image of x_i is the core circle of T_i. So the resultant manifold still projects to F. The inverse image of each point in F is a circle (called a *fiber*). The inverse image of each x_i is called an *exceptional fiber* and other fibers are called *regular fibers*.

The description of the Dehn surgery is not yet complete since there is still choice in how the cross-section $\partial E_i \times \{y\}$ is attached to ∂T_i. Any first choice could be altered by Dehn twists, in ∂T_i, along the (p_i, q_i) torus knot image of $\{x\} \times S^1$, so the possible choices are parameterized by the integers. But there is also another ambiguity. There may be automorphisms of $F_- \times S^1$ which preserve the fiber structure (i.e., commute with projection to F_-). It is easy to see that, when F has boundary, all of the choice of cross-section attachment can be absorbed into the ambiguity of what is the global cross-section $F_- \times \{y\}$, so the data above is sufficient to characterize the manifold up to homeomorphism that commutes with projection to F. If F is closed, there is not so much flexibility of the choice of cross Section so ultimately there is an integer's worth of choice involved in how the manifold is constructed. Typically this choice is realized by choosing another point $x_0 \in F_-$ and disk $E_0 \subset F_-$ containing it and doing Dehn surgery on $E_0 \times S^1$, gluing in a solid torus T_0 so that its longitude is identified with $\{x\} \times S^1 \subset \partial E_0 \times S^1$ and its meridian is identified with a cross-section of $\partial E_0 \times S^1$, of which there are an integer's worth of possibilities. The choice here determines what is called the *Euler number* of the Seifert manifold. (Working out the details in this paragraph is a good first step at understanding obstruction theory.)

There are two ways in which the Seifert structure can induce natural Heegaard splittings on the 3-manifold and these are the subject of the next two sections. It is the principal result of [32] that any irreducible splitting of a fully orientable Seifert manifold is one of these two types.

4.1. *Vertical splittings*

Suppose that M is a fully orientable Seifert manifold, constructed as above, with base surface F, projection $p: M \to F$, and singular fibers the inverse images of $x_1, \ldots, x_n \in F$. Let Γ be a connected graph in F_- chosen so that

(i) some nonempty subset of the x_i, $1 \leqslant i \leqslant n$, are vertices of Γ;

(ii) each component of $F - \Gamma$ is either a disk containing a single x_i or an annulus containing a single boundary component. (But if F is closed and $n = 1$, then $F - \Gamma$ is a disk not containing x_1.)

(iii) $\Gamma \cap \partial F$ consists of a (possibly empty) collection of boundary components $d_1, \ldots, d_d$.

Let $H_1 \subset M$ be the compression body whose spine is the union of $\{d_j \times S^1, j = 1, \ldots, d\}$, a lift of Γ, and the singular fiber lying over each $x_i \subset \Gamma$. The complement of H_1 in M is also a compression body, whose spine is the union of the boundary components of M not in $\{d_j \times S^1\}$, the exceptional fibers not lying over Γ, and the lift of a "dual" complex to Γ. This creates a Heegaard splitting which is called a *vertical* Heegaard splitting.

It is not difficult to show that, up to isotopy, this construction is independent of Γ but depends only on the choice of the x_i that lie in Γ and the choice of the boundary components $d_1, \ldots, d_d$.

THEOREM 4.1 ([45]). *If M is a fully orientable Seifert manifold and $\partial M \neq \emptyset$ then any irreducible Heegaard splitting is vertical.*

PROOF. Here is a sketch of the complicated and ingenious argument. The proof is by induction on the number of exceptional fibers, with the case of no such fibers (i.e., $M = F \times S^1$) covered in [44]. Let e be an exceptional fiber, put in "thin" position with respect to a sweep-out (Section 2.4) coming from the Heegaard splitting. This means, roughly, that if one considers how the circle fiber e is intersected by the sweep-out, one cannot move the levels at which the maxima and minima of e occur by pushing a maximum down and a minimum up until the the maximum is encountered before the minimum. This so simplifies e that it can be moved to lie on a Heegaard surface in a way so that it intersects a meridian curve on one side in a single point. This is sufficient to ensure that e can be made a core of a handle on one side, so that removing it from M leaves the Heegaard surface as the splitting surface of $M - \eta(e)$. □

4.2. *Horizontal splittings*

Here is a specialized way to construct some fully orientable Seifert manifolds. Let $\widehat{F}$ be an orientable compact surface and $h : \widehat{F} \to \widehat{F}$ be a periodic orientation preserving diffeomorphism, such that $h^n = identity$. Consider the mapping cylinder of h, a compact 3-manifold M obtained by identifying, in $\widehat{F} \times I$, each point $x \times 0 \in \widehat{F} \times \{0\}$ with $h(x) \times 1$. Notice that M is fibered by circles. For any point $x \in \widehat{F}$ the union of the images of $\{h^i(x)\} \times I \subset \widehat{F} \times I$, $1 \leqslant i \leqslant n$, is a circle which typically intersects a cross-section $\widehat{F} \times \{s\}$ in n points. For some discrete (hence finite) set of points in $\widehat{F}$, the orbit may be of length only a proper factor l of n and then the corresponding circle intersects a cross-section only in l points. It's easy to see that this gives M a Seifert manifold structure in which the base surface is $F = \widehat{F}/h$, a surface over which $\widehat{F}$ is a branched covering.

A Seifert manifold can be given such a structure if and only if its Euler number is zero (see [47]). In particular, any fully orientable Seifert manifold with non-empty boundary can be given such a structure. Suppose M_- is a fully orientable Seifert manifold whose boundary is a single torus T. View M_- as the mapping cylinder of a diffeomorphism $h : \widehat{F} \to \widehat{F}$, where $\widehat{F}$ is a compact orientable surface with a single boundary component. Now $\partial \widehat{F} \subset \partial M_-$ is a circle c transverse to the fibers of M_-. Attach a solid torus T to ∂M_- so that a longitude goes to c. (There is an integer's worth of choice of how the meridian is attached.) This creates a closed Seifert manifold M which can be split into two pieces: The image of $\widehat{F} \times [0, 1/2]$ and the union of T and $\widehat{F} \times [1/2, 1]$. Since both pieces are homeomorphic to $\widehat{F} \times I$ (in the latter case because T becomes just a collar of $\partial \widehat{F} \times I$), each is a handlebody. Thus we get a Heegaard splitting, and this construction is called a *horizontal* splitting of M.

THEOREM 4.2 ([32,46]). *An irreducible Heegaard splitting of a fully orientable Seifert manifold is either horizontal or vertical.*

PROOF. A special argument [46] is needed for small Seifert manifolds; we sketch here the proof when M contains an essential vertical torus. Suppose $H_1 \cup_S H_2$ is the irreducible splitting and suppose that it is weakly reducible. Since the splitting is irreducible it follows from the proof of Theorem 3.11 that, if S is maximally and independently compressed in both directions, the result is an incompressible surface $\bar{S}$. Furthermore S can be viewed as assembled from Heegaard splittings of $M - \bar{S}$ by amalgamating along $\bar{S}$. Any incompressible surface in a Seifert manifold can be isotoped to be either a collection of vertical tori, or the fiber in a fibering of M over S^1.

Suppose $\bar{S}$ is a collection of vertical tori. Then $M - \bar{S}$ is a Seifert manifold with boundary, and so any Heegaard splitting is vertical, by Theorem 4.1. Thus S is obtained by amalgamating vertical splittings along vertical tori, from which it follows immediately that S is also vertical.

Suppose $\bar{S}$ is a set of fibers of a fibering over S^1. Then it splits M into pieces of the form $\bar{S} \times I$, and induces a Heegaard splitting on each piece. Heegaard splittings of *surface* $\times I$ are well-understood [41] and examination shows that in fact the compressing could have been done in such a way that $\bar{S}$ would be a collection of vertical tori, reducing to the previous case.

Suppose finally that S is strongly irreducible. An argument similar in spirit to the thin position argument of Theorem 4.1 proves that a fiber f can be isotoped onto the surface S. Then the Seifert manifold $M_- = M - \eta(f)$ is split in two pieces by the surface $S_- = S - \eta(f)$. If S_- is incompressible in M_- then it is the fiber of a fibration of S_- over S^1 and it follows easily that the original S is a horizontal splitting. If S_- is compressible then, since S is strongly incompressible, after a maximal number of compressions into one handlebody, say H_2, S_- becomes an incompressible surface $S*$ in M_-.

If S^* is a vertical annulus then the union of S^* and $S \cap \eta(f)$ is a torus in H_1 so it bounds a solid torus. The core of the torus is a fiber and the manifold obtained by deleting it has S as a Heegaard splitting surface. It follows from 4.1 that S is vertical.

If S^* is the fiber of a fibering of M_- over S^1 then S^* splits M_- into handlebodies and S is a further Heegaard splitting of one them. But any (non-trivial) splitting of a handlebody is stabilized, hence reducible (from Proposition 3.6 and Theorem 3.7) and S is assumed irreducible. □

On the other hand, not all vertical and horizontal splittings are irreducible. Exactly which ones are has been worked out in [49].

5. Connections with group presentations

In this section, assume that $M = H_1 \cup_S H_2$ is a closed manifold, and hence that both H_1 and H_2 are handlebodies, say of genus g. This implies that $\pi_1(H_1)$ is a free group on g generators. A choice of base-point and a complete collection $\Delta = \{D_1, \ldots, D_g\}$ of oriented meridian disks determines a presentation of $\pi_1(H_1)$, namely, for any based loop in H_1, write down x_i every time the loop passes through the disk D_i in a direction consistent with its normal orientation and x_i^{-1} if the direction is inconsistent. Similarly, a complete collection $E_1, \ldots, E_s$, $s \geqslant g$, of meridian disks for H_2 then determines a presentation of $\pi_1(M)$. Each curve ∂E_k, when viewed as a (conjugacy class) in $\pi_1(H_1)$, and so as a word r_k in $\{x_i\}$, is a relator for the fundamental group. That is, $\pi_1(M)$ has the presentation $\{x_1, \ldots, x_g; r_1, \ldots, r_s\}$. We say that this presentation is *geometrically realized*.

How much does this presentation depend on choices made? We'll restrict attention to H_1 (which yields the generators) since the situation is much the same for the relators.

LEMMA 5.1. *Any set $\{y_1, \ldots, y_g\}$ of generators of $\pi_1(H_1)$ can be geometrically realized.*

PROOF. There are a specific set of moves on generators, called the Nielsen moves, which will transform a given geometrically realized set of generators $\{x_1, \ldots, x_g\}$ into $\{y_1, \ldots, y_g\}$. But an examination of these moves (see, e.g., [27, 3.1]) shows that each move can be realized by a geometric move, either sliding one 2-handle over the other (i.e., band-summing one meridian disk to another) or reversing the orientation of a disk, or just naming the disks in a different order. □

Motivated in part by this lemma it makes sense to introduce the following definition.

DEFINITION 5.2 ([26]). Two generating systems $\{u_1, \ldots, u_g\}$ and $\{v_1, \ldots, v_g\}$ for $\pi_1(M)$ are *Nielsen equivalent* if there is an epimorphism $\phi: F_g \to \pi_1(M)$ and bases $\{x_1, \ldots, x_g\}$ and $\{y_1, \ldots, y_g\}$ for the free group F_g, such that $\phi(x_i) = u_i$ and $\phi(y_i) = v_i$.

Less formally, if we view a presentation of G as an epimorphism of the free group (with specified generators) onto G, then two presentations are Nielsen equivalent if there is an automorphism of the free group which realizes the change in specified generators.

It follows that a Heegaard splitting $M = H_1 \cup_S H_2$ (and a choice of which handlebody is H_1) specifies a single Nielsen equivalence class of presentations of $\pi_1(M)$. For if $\{u_1, \ldots, u_g\}$ and $\{v_1, \ldots, v_g\}$ are the generating systems induced by different choices of meridian disks for H_1, then in the definition above substitute $\pi_1(H_1)$ for F_g, let the inclusion induce ϕ, and deduce that the presentations are Nielsen equivalent.

To generalize slightly:

THEOREM 5.3. *If two Heegaard splittings $H_1 \cup_S H_2$ and $H_1' \cup_{S'} H_2'$ of the same closed 3-manifold M are isotopic then (for the appropriate choice of H_1 and H_1') their corresponding geometrically realized presentations are Nielsen equivalent.*

PROOF. Inner automorphism is a Nielsen equivalence. □

This gives a powerful algebraic tool to show that Heegaard splittings are not isotopic.

The structures of Heegaard splittings discussed in Section 3 above have implications for the induced group presentations. For example, if a Heegaard splitting is stabilized, then there is a Nielsen equivalent presentation in which a relator is precisely a generator. If it is reducible, then there is a Nielsen equivalent presentation which splits as a free product of two presentations. (The genus one splitting of $S^1 \times S^2$ and also the non-orientable 2-sphere bundle over S^1 are exceptions. And of course the groups presented might be trivial.) If it is weakly reducible, then there is a Nielsen equivalent presentation in which at least one generator does not appear in at least one relator.

Not all presentations can be geometrically realized. For example, Boileau and Zieschang [8] have shown that certain Seifert manifolds have fundamental groups which admit presentations with two generators, whereas the minimal genus of any Heegaard splitting (hence the rank of any geometrically realized presentation) is at least three. Montesinos [30] has used this example to show that a presentation Nielsen equivalent to a geometric presentation may not be geometric.

For details on this and other examples, see [53].

6. Uniqueness

How many distinct Heegaard splittings does a 3-manifold have? We have already seen that any Heegaard splitting can be stabilized, so the question only becomes interesting if we restrict to Heegaard splittings which are not stabilized.

6.1. *The 3-sphere*

In 1968 Waldhausen [52] showed that any positive genus Heegaard splitting of S^3 is stabilized (3.7), so that the only genus g splitting is the obvious one, obtained by stabilizing q times the splitting of S^3 into 3-balls. This was the first uniqueness result. Here is a sketch of a later proof [42,33].

THEOREM 3.7. *Any positive genus Heegaard splitting of S^3 is stabilized.*

PROOF. Suppose $S^3 = H_1 \cup_S H_2$ and Σ is a spine of H_1. We may assume Σ is a trivalent graph in S^3 and we are allowed to do edge-slides. Choose a Morse function $h : S^3 \to [-1, 1]$ which has a single minimum (at height -1) and a single maximum (at height 1) and which restricts to a Morse function on Σ. Put Σ in "thin position" with respect to this height function. In outline, this means that you can't push down a maximum (this includes trivalent vertices in which two edges leave the vertex from below) so that it moves below a minimum (this includes trivalent vertices in which two edges leave the vertex from above) without introducing new critical points.

It suffices to show there is an unknotted cycle $\gamma \subset \Sigma$. For then S would also be a Heegaard splitting surface for the solid torus $S^3 - \eta(\gamma)$. This splitting would necessarily be boundary reducible (Proposition 3.6) which means that the original splitting S was stabilized.

Consider a collection $\Delta \subset S^3$ of meridian disks of H_2, extended into H_1, so that its interior is embedded in $S^3 - \Sigma$ and its (singular) boundary lies in Σ. The first observation is that we may as well assume $\partial \Delta$ runs across every edge of Σ, for otherwise $H_1 \cup_S H_2$ would be reducible (Theorem 3.11). If the splitting were reducible then a reducing sphere splits S into two Heegaard splittings of S^3 each of smaller positive genus, and we would be done by induction.

Consider when a level sphere $S_t = h^{-1}(t)$ cuts off from Δ a subdisk sufficiently simple that it can be used to slide part of an edge of Σ so that it lies on S_t. It's easy to see that this is true just below the highest point of Σ and just above the lowest point. In the former case the disk can be used to lower the maximum slightly and in the latter to raise the minimum. Suppose we simultaneously (i.e., for the same level sphere) have two subdisks of Δ, one of which lowers a maximum and the other of which raises a minimum. Then either this violates thin position (when we can push the maximum slightly lower without interfering with the minimum) or the two edges which we have pushed onto the level sphere have the same ends, i.e., they create an unknotted cycle and we are done.

We know then that a sufficiently high sphere cuts off a subdisk of Δ lowering a maximum, a sufficiently low sphere cuts off a subdisk raising a minimum and, if subdisks of both types are cut off simultaneously, then we are done. So it suffices to eliminate the possibility that neither type occurs, that is, there is a height t_0 so that no subdisk cut off by S_{t_0} from Δ can be used either to raise a minimum or lower a maximum. But this situation cannot in fact occur, by an argument reminiscent of the second proof of Theorem 3.4, with S_{t_0} playing the role of the reducing sphere. □

6.2. *Seifert manifolds*

One might have hoped that this situation would generalize – that any compact 3-manifold would have (up to stabilization) a unique Heegaard splitting. In 1970 Engmann ([15], see also [4]) showed that the connected sum of certain pairs of Lens spaces could have two non-homeomorphic Heegaard splittings of genus two (hence not stabilized). Examples were shortly found of prime manifolds with the same property [5]. A rather spectacular generalization by Lustig and Moriah is the main theorem of [26]. It is a good illustration of the usefulness of Theorem 5.3, so we sketch the central idea.

Let M be a fully orientable Seifert manifold, constructed as in Section 4, with base surface F, projection $p: M \to F$, and singular fibers the inverse images of $x_1, \ldots, x_n \in F$. Given details of the fibering around the x_i and the Euler number of M it is straightforward to write down a presentation of $\pi_1(M)$. It's easy to see directly that the element $h \in \pi_1(M)$ represented by a regular fiber is central in $\pi_1(M)$.

Consider the quotient group $G = \pi_1(M)/\langle h\rangle$. The complement M_- of the exceptional fibers is $F_- \times S^1$ so the effect of factoring out $\langle h\rangle$ is to reduce $\pi_1(M_-)$ to $\pi_1(F_-)$. In the solid torus surrounding an exceptional fiber a meridian crosses a fiber some $p_i \geqslant 2$ times. The effect is to kill the p_i multiple of ∂E_i in $\pi_1(F_-)$. The upshot is that G is a Fuchsian group and in particular has a faithful presentation into $PSL_2(\mathbf{C})$. This special structure provides an extra tool for determining when group presentations are Nielsen equivalent. (Note that Nielsen equivalent presentations of $\pi_1(M)$ descend to Nielsen equivalent presentations of G.)

This extra information is sufficient to show that, in most cases, two vertical splittings of the same fully orientable 3-manifold are isotopic only if the equivalence is more or less obvious, e.g., the invariants of the exceptional fibers lying in the graph Γ are the same (see Section 4.1). In particular this leads to a complete classification of irreducible Heegaard splittings of most fully orientable Seifert 3-manifolds with boundary (see [45]). In the case of closed Seifert manifolds, there is still some puzzlement about how horizontal splittings fit into the classification scheme. For example, whereas a vertical splitting of a closed Seifert manifold with base surface of genus g and with k exceptional fibers is $2g + k - 1$, there are some such Seifert manifolds (over S^2, i.e., $g = 0$) which have horizontal splittings of genus $k - 2$.

6.3. *Genus and the Casson–Gordon examples*

The last comment prompts the following question: Do we at least know that all irreducible splittings of the same 3-manifold have the same genus? In 1986 Casson and Gordon gave an example which shows that the answer is an emphatic no [13,25]. What they show is that there is a closed orientable 3-manifold (in fact infinitely many) which has irreducible splittings of arbitrarily high genus. We outline the construction.

Begin with the following fact [34]: There are certain pretzel knots $k \subset S^3$ with the property that they have incompressible Seifert surfaces of arbitrarily high genus (these are explicitly constructed) and for each of these surfaces the complement in S^3 is a handlebody. Pick one of these knots, and let F_n be an incompressible Seifert surface of genus n whose complement in S^3 is a handlebody. Then $S^3 = \eta(F_n) \cup (S^3 - int(\eta(F_n)))$ is a (highly reducible) genus $2n$ Heegaard splitting of S^3, and k is isotopic to a curve on the splitting surface $S = \partial\eta(F_n)$. Let M_q be the 3-manifold obtained by doing $1/q$ surgery on k (q an integer). One way to view this is to imagine pulling the two handlebodies apart along a strip parallel to $k \subset S$ then gluing the two strips back together via a q-fold Dehn twist. So in particular the construction naturally gives a genus $2n$ Heegaard splitting of M_q. For q a large integer ($q \geqslant 6$ suffices) it turns out (see below) that the resulting splitting is strongly irreducible. Thus a specific M_q will have splittings, built as above for different values of n, of arbitrarily high genus.

The critical ingredient in the above argument is then

THEOREM 6.1 ([13]). *Suppose $M = H_1 \cup_S H_2$ is a weakly reducible Heegaard splitting of the closed manifold M. Let k be a simple closed curve on the splitting surface S so that $S - \eta(k)$ is incompressible in both H_i, $i = 1, 2$. Let M_q be the manifold obtained by $1/q$ surgery on k. Then for $q \geqslant 6$ the associated Heegaard splitting (induced as above) on M_q is strongly irreducible.*

See [32, Appendix] for a proof. The idea is this: k necessarily intersects any meridian disk on either side, since $S - \eta(k)$ is incompressible on both sides. Sufficient Dehn twisting along k then will stretch any meridian of one side so that it intersects any meridian disk of the other.

In fact [25] the number of Heegaard splittings at each even genus is bounded below by a polynomial in the genus.

6.4. *Other uniqueness results*

We briefly note that there are other manifolds which are known to have unique irreducible Heegaard splittings (for a particular distribution of boundary components between H_1 and H_2). A perhaps not exhaustive list is the following:
- $S^2 \times S^1$ [52];
- Any Lens space [9,10] (see also 7.8);
- Any *(closed orientable surface)* $\times I$ [41,7];
- Any *(compact orientable surface)* $\times S^1$ [44,7].

7. The stabilization problem

We noted in Section 2.1 that every compact 3-manifold admits a Heegaard splitting, since every 3-manifold has a triangulation. Similarly, since any two triangulations of the same 3-manifold are PL equivalent (see [29,3]) it follows that any two Heegaard splittings have a common stabilization. The argument, which goes back to Reidemeister and Singer, is more complicated than one might expect. See [1] for details.

On the other hand, we noted in Section 6.2 that the connection between Heegaard splittings and group presentations and the known structure of Heegaard splittings of Seifert manifolds show that some manifolds have distinct irreducible Heegaard splittings. This raises the natural question: How much do we need to stabilize before two splittings of the same 3-manifold become isotopic? More generally, now that we know that a manifold can have quite different Heegaard splittings, how can such distinct Heegaard splittings be compared?

As a cautionary tale, revealing the depth of our ignorance on the first question, consider the large gap between what is known and what is not:

THEOREM 7.1 ([46]). *Two irreducible Heegaard splittings of the same fully orientable Seifert manifold have a common stabilization requiring, for one of the splittings, at most one stabilization.*

THEOREM 7.2 ([48]). *For the irreducible Heegaard splittings of M_q constructed in Section 6.3, any two Heegaard splittings of the same M_q have a common stabilization requiring, for one of the splittings, at most one stabilization.*

In fact, there is no example of distinct Heegaard splittings of the same closed 3-manifold which cannot be made isotopic by a single stabilization of one of the splittings, and sufficient stabilizations of the other to ensure that the genus of the two surfaces is the same. One could thus make the very optimistic

CONJECTURE 1. Suppose $H_1 \cup_S H_2$ and $H_1' \cup_{S'} H_2'$ are Heegaard splittings of the same 3-manifold of, genus $g \leqslant g'$ respectively. Then the splittings obtained by one stabilization of S' and $g' - g + 1$ stabilizations of S are isotopic.

At the other extreme are two theorems which put limits on how much stabilization is needed, in terms of the genera of the two original splittings.

THEOREM 7.3 ([24, Theorem 31.9]). *Suppose M is a Haken 3-manifold containing no non-trivial essential Stallings fibrations. Then the number of stabilizations required to guarantee that a genus g splitting of M is isotopic to a genus g' splitting is some polynomial function (perhaps depending on M) of g and g'.*

The gap here is rather huge. An ideal sort of theorem would be one which gives an explicit bound, independent of the manifold, on the number of stabilizations required, expressed in terms of the genera of the splittings being considered. This would be an important step toward solving the "homeomorphism problem" – find an algorithm which will determine if two compact 3-manifolds are homeomorphic – because it would reduce the problem to the case in which there are isotopic Heegaard splittings of the same known genus for the two manifolds.

In this direction is the following theorem, which applies to all irreducible splittings of compact orientable non-Haken 3-manifolds:

THEOREM 7.4 ([38, Theorem 11.5]). *Suppose $X \cup_Q Y$ and $A \cup_P B$ are strongly irreducible Heegaard splittings of the same closed orientable 3-manifold M and are of genus $p \leqslant q$ respectively. Then there is a genus $8q + 5p - 9$ Heegaard splitting of M which stabilizes both $A \cup_P B$ and $X \cup_Q Y$.*

It appears that a similar explicit, but quadratic, bound can be found for Haken 3-manifolds using [40] and [39]. The former extends Theorem 7.4 to the bounded case. The machinery of the latter, [39], illustrates how to extend certain general position arguments of [38] to weakly reducible splittings that have been untelescoped, as described in Section 3.9.

The proof of Theorem 7.4 is quite complicated, but a crucial ingredient is a theorem that describes how two strongly irreducible splittings can be moved to intersect in a way that contains much information about both splittings.

THEOREM 7.5 ([38, Theorem 6.2]). *Suppose $X \cup_Q Y$ and $A \cup_P B$ are strongly irreducible Heegaard splittings of the same closed orientable 3-manifold $M \neq S^3$. Then P and Q can be isotoped so that $P \cap Q$ is a non-empty collection of curves which are essential in both P and Q.*

The proof is an application of Heegaard splittings as sweep-outs (Section 2.4). For $A \cup_P B$ a Heegaard splitting of a closed 3-manifold M, let Σ_A and Σ_B be spines of the handlebodies A and B respectively. Recall

DEFINITION 7.6. A *sweep-out* associated to the Heegaard splitting $A \cup_P B$ is a relative homeomorphism $H : P \times (I, \partial I) \to (M, \Sigma_A \cup \Sigma_B)$ which, near $P \times \partial I$, gives a mapping cylinder structure to a neighborhood of $\Sigma_A \cup \Sigma_B$.

Given such a sweep-out H and a value s, $0 \leqslant s \leqslant 1$, let P_s denote $H(P \times s)$, $P_{<s}$ denote the handlebody $H(P \times [0, s])$ and $P_{>s}$ denote the handlebody $H(P \times [s, 1])$. Note that $P(s)$, $s \neq 0, 1$, is a copy of the splitting surface, $P_0 = \Sigma_A$ and $P_1 = \Sigma_B$.

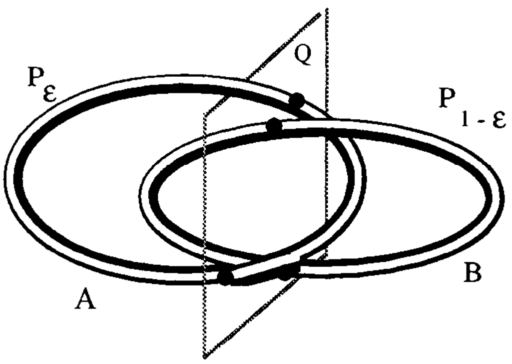


Fig. 8.

Consider how the surfaces P_s intersect a distinct Heegaard splitting surface Q in $M = X \cup_Q Y$. Assume Q is in general position with respect to $\Sigma_A \cup \Sigma_B$ (so the spines intersect Q transversally in a finite number of points) and the sweep-out H is generic with respect to Q. Then, for small values of ε, $P_{<\varepsilon}$ is very near Σ_A, so $P_{<\varepsilon} \cap Q$ is a (possibly empty) collection of meridian disks of A. Symmetrically $P_{>1-\varepsilon}$ is very near Σ_B, so $P_{>1-\varepsilon} \cap Q$ is a (possibly empty) collection of meridian disks of B. Throughout the sweep-out, at least generically, $P_s \cap Q$ is a disjoint collection of simple closed curves in Q. (See Figure 8.)

Note that $P_s \cap Q$ cuts off in Q meridian disks for A when s is small, meridian disks for B when s is large and can't cut off simultaneously meridian disks for both, since $A \cup_P B$ is strongly irreducible. It follows that for some value of s, no meridian is cut off. That is (with a minor amount of additional fuss) every curve of $P_s \cap Q$ is essential in Q.

In order to prove Theorem 7.5 we would like to apply a similar argument simultaneously to sweepouts P_s, Q_t of M corresponding to the different Heegaard splittings of M. Cerf theory (see [14]) can be used to make the following informal remarks rigorous. A good way to think visually of the discussion below is to consider the surfaces P_s and Q_t as parameterized by the point (s, t) in the square $I \times I = \{(s, t) \mid 0 \leqslant s, t \leqslant 1\}$.

Away from $\partial(I \times I)$, four things can happen:

- At a generic value of (s, t), P_s and Q_t intersect transversally in a collection of simple closed curves $c_{(s,t)}$ which we can regard as lying in either $P \cong P_s$ or $Q \cong Q_t$.
- On a one-dimensional stratum of $I \times I$, P_s and Q_t intersect transversally except at a single non-degenerate tangency point. A good way to think about this is to begin at a value of (s, t) at which there is such a tangency point. Now imagine letting s ascend (or descend) at just the rate required to ensure that the tangency point persists as t ascends. This requirement defines s as a function of t, and so parameterizes an arc inside the square $I \times I$. (Note that the slope of the arc is positive or negative depending on whether the ascending normal vectors to P_s and Q_t are parallel or anti-parallel. Thus the sign of the slope is fixed, providing a surprising order to the picture. So far, this additional order has not proven useful.)
- A discrete set of points (s, t) for which P_s and Q_t have exactly two non-degenerate points of tangency but are otherwise transverse. For example, as (s, t) traces out the arc

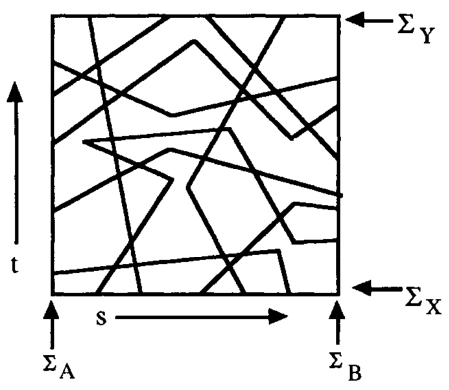


Fig. 9.

as just described, there may be points of tangency which occur elsewhere. These are the discrete critical points of double tangency.

- A discrete set of points at which P_s and Q_t intersect transversally except for a single degenerate tangent point (locally modelled on $P_s = \{(x, y, z) \mid z = 0\}$ and $Q_t = \{(x, y, z) \mid z = x^2 + y^3\}$). These are so-called "birth-death" points, and play no important role in our discussion.

The set of points (s, t) at which the intersection is non-generic forms a 1-complex Γ called the *graphic* of the sweep-outs in the interior of $I \times I$. The graphic Γ naturally extends to a properly imbedded 1-complex in all of $I \times I$: A point $(0, t)$, say, on $\{0\} \times I \subset \partial(I \times I)$ represents simultaneously the spine Σ_A of handlebody A (since $s = 0$) and the surface Q_t. Generically these are transverse, implying that P_ε and Q_t are transverse for ε small. There are two types of exceptions: For finitely many values of t, Σ_A is tangent to Q_t at a single point in the interior of one of its edges. At finitely many other values of t, Q_t crosses a vertex of Σ_A. It's easy to see that, in both these cases, nearby interior points are in the graphic, and vice versa, so Γ extends to a graphic in the closed square. (See Figure 9.)

PROOF OF THEOREM 7.4. The graphic Γ cuts $I \times I$ up into regions, in each of which the curves $c_{(s,t)}$ vary only by an isotopy in P and Q. In some regions (e.g., near $\{0\} \times I$), $c_{(s,t)}$ cuts off meridian disks for A lying in Q. In other regions (e.g., near $\{1\} \times I$), $c_{(s,t)}$ cuts off meridian disks for B lying in Q. One can't have both occur in the same region, since P is strongly irreducible. (Indeed they can't even occur in adjacent regions, but this is not immediately obvious.) Similarly there are regions in which $c_{(s,t)}$ cuts off a meridian of X or Y, but not both, in P. We now apply a "mountain-pass" sort of argument: Given what we have described, there must be some point in the interior of $I \times I$ in which $c_{(s,t)}$ cuts off no meridians whatsoever. Such a point is the point we seek. It corresponds to an intersection in which $c_{(s,t)}$ is non-empty and each curve is essential in both P and Q. (See Figure 10.)

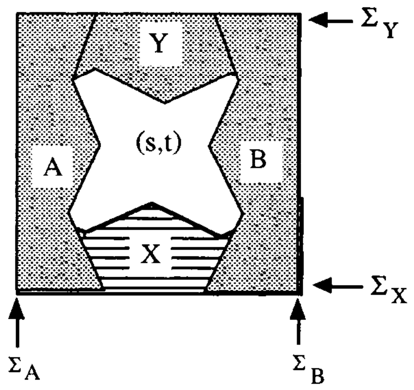


Fig. 10.

What is not apparent in the above argument is why the point (s, t) we have located is a generic point, nor is it clear how we can guarantee that the intersection $c_{(s,t)}$ is non-empty at this point. The details here require close argument, see [38]. □

To illustrate the power of this argument, we classify the irreducible splittings of the Lens space [9,10]. But first observe that (with a bit of reorganization) the argument above shows that any positive genus Heegaard splitting of S^3 is stabilized: Compare such a splitting with the index zero splitting (i.e., by S^2) of S^3. How could a curve of intersection with S^2 be essential in S^2?

COROLLARY 7.7 ([9]). *Any two genus one Heegaard surfaces in a lens space are isotopic.*

PROOF. Let P and Q be two genus one Heegaard surfaces in a lens space, separating the lens space, as usual, into solid tori A and B and solid tori X and Y respectively. P and Q may be isotoped so that they intersect in a non-empty family of essential circles, at which point $c_{(s,t)}$ cuts each up into annuli. One can pass annuli of P parallel to annuli of Q through each other until only two curves of intersection remain. At this point it is easy to show that the remaining annuli of $P - Q$ are parallel to the annuli $Q - P$. This means P is parallel to Q. □

COROLLARY 7.8 ([10]). *Any irreducible Heegaard splitting of a lens space has genus one.*

PROOF. Let $A \cup_P B$ be a genus one Heegaard splitting of a lens space L and $X \cup_Q Y$ be a splitting of higher genus. Since L contains no incompressible surfaces, it suffices to show that Q is weakly reducible. P and Q may be isotoped so that they intersect in a non-empty family of essential circles. There are two cases:

Case 1. $Q \cap A$ and $Q \cap B$ both contain components which aren't annuli.

As above, remove parallel annuli in P and Q by an isotopy. Then one can show, by ∂-compressing Q in both A and B, that somewhere in A there is a meridian of X and somewhere in B a meridian of Y (or vice versa). This contradict the strong irreducibility of Q. So we are reduced to

Case 2. $Q \cap A$ or $Q \cap B$ (say the former) consists entirely of annuli.

When we remove parallel annuli, one can show that in the end, Q actually lies in the solid torus B and in fact induces a Heegaard splitting of B. But it follows from Proposition 3.6 that any higher genus Heegaard splitting of a solid torus is stabilized. □

8. Normal surfaces and decision problems

8.1. *Normal surfaces and Heegaard splittings*

Much of our understanding of 3-manifolds depends on the surfaces they contain. Their most elementary taxonomy is expressed by reference to these surfaces: M is irreducible if it contains no essential sphere, it's Haken if it contains a higher genus incompressible surface, it's atoroidal (and so, if closed and Haken, hyperbolic) if no such incompressible surface is a torus. Heegaard splittings have proven fundamental both in understanding the behavior of these surfaces and in developing algorithms (typically impractical) for classifying 3-manifolds within this taxonomy.

We briefly review the theory of normal surfaces [16]. A good source for more detail is [23].

Let $M = H_1 \cup_S H_2$ be a Heegaard splitting of a closed manifold (the case where M is merely compact is an easy variation). Regard each H_i as a handlebody, the union of some 0-handles (one would suffice) and some 1-handles. Suppose F is a closed surface in M. Then F can be isotoped so that it is disjoint from the points which are the cores of the 0-handles of H_2 and intersects transversally each of the cores of the 1-handles. By thickening these cores to the full handle-body we can isotope F so that it intersects H_2 in some finite number of copies of 2-disk cocores Δ_2 (meridians) of the 1-handles of H_2. Call the number of such disks in F the *weight* of F.

Consider how F then intersects H_1. It is helpful to recall the discussion of Heegaard diagrams in Section 2.3. The handlebody H_1, when cut up by a family Δ_1 of meridians, becomes a collection of 3-balls, the 0-handles of H_i. With little loss of generality we will assume in this discussion that there is a single 3-ball, B^3. In ∂B^3, the attaching curves of Δ_2 become a 1-manifold $\mathcal{A}$ in $\partial B^3 - V$. The arcs are regarded as edges of a graph Γ whose vertices V correspond two-to-one to the meridians Δ_1 of H_1. (We will here expand Γ to include the simple closed curves of $\mathcal{A}$.) We may as well assume that F is transverse to ∂B^3, so that $F \cap B^3$ is a properly imbedded surface lying in B^3. Because we have already assured that $F \cap H_2$ consists of copies of Δ_2 we know that the collection of simple closed curves $F \cap \partial B^3$ is the union of parallel copies of components of $\mathcal{A}$ outside of V and, inside of V, consists of some properly imbedded 1-manifold.

DEFINITION 8.1. The surface F is *normal* with respect to $H_1 \cup_S H_2$ if

(i) Each component of $F \cap B^3$ is a disk.

(ii) No component of $F \cap \partial B^3$ lies entirely in a fat vertex v.

(iii) Each component of $F \cap \partial B^3$ contains at most one copy of any edge of Γ.

DEFINITION 8.2. A property of surfaces in 3-manifolds is called *compression preserved* if whenever a surface F in M has this property, and F' is obtained from F by a 2-surgery, then some set of components of F' (not inessential spheres) also has this property.

Examples of compression preserved properties are

- F is a reducing sphere.
- F is an injective surface (i.e., $\pi_1(F) \to \pi_1(M)$ is injective).
- F has maximal Euler characteristic in its homology class (ignoring inessential spheres).

We then have:

THEOREM 8.3. *If a closed 3-manifold M contains a surface with a compression preserved property, then it contains a normal surface with the same compression preserved property.*

PROOF. Choose a surface in M which has the property and also has minimal weight. By compressing along disks lying in V we can remove any components of $F \cap \partial B^3$ that lie completely in V. By compressing along disks lying slightly inside ∂B^3 we can arrange that the surface intersects B^3 in disks and in components lying entirely inside B^3. The latter can be discarded since they compress to inessential spheres and so, by Definition 8.2, they do not have the property. Finally, if any component of $F \cap \partial B^3$ contains more than one arc parallel to an edge γ in Γ then there is a ∂-compression of the corresponding disk in $F \cap B^3$ to an arc in $\eta(\gamma) \subset \partial B^3$. Then push across a 2-handle, reducing the weight of F by two, a contradiction. □

Since there are only a finite number of edges (and simple closed curves) in Γ, there are only a finite number of isotopy types of simple closed curves in $\eta(\Gamma) \subset \partial B^3$ which can arise as components of $F \cap \partial B^3$. Thus any normal surface can be described completely by saying how many of each of the finite number of possible types occur. Normal surfaces are useful in constructing algorithms because the decisions made in creating a normal surface are essentially finite. The theory is even more powerful than these considerations suggest, since the operation of adding the numbers that classify two surfaces has geometric content. For a full appreciation, it is helpful to consider the very specific type of Heegaard splitting that comes from a triangulation.

8.2. *Special case: Normal surfaces in a triangulation*

Let M be a closed triangulated 3-manifold with a fixed triangulation $\mathcal{T}$. Let T^i denote the i-skeleton of $\mathcal{T}$. We will consider what it means for a surface to be normal in the induced Heegaard splitting $H_1 \cup_S H_2$ where, in contrast to 2.1, H_2 is a neighborhood of T^1 and H_1 is a neighborhood of the dual 1-skeleton.

Suppose F is a closed surface in M. Then the requirement in 8.1 that F be disjoint from the 0-handles in H_2 and intersect the cores of the 1-handles of H_2 and ∂B^3 transversally here translates to the requirement that F be in general position with respect to the triangulation $\mathcal{T}$. The weight of F is just the number $|F \cap T^1|$. Since H_1 is a neighborhood of the dual complex to the triangulation, the 2-simplices of $\mathcal{T}$ are a collection Δ_1 of meridians

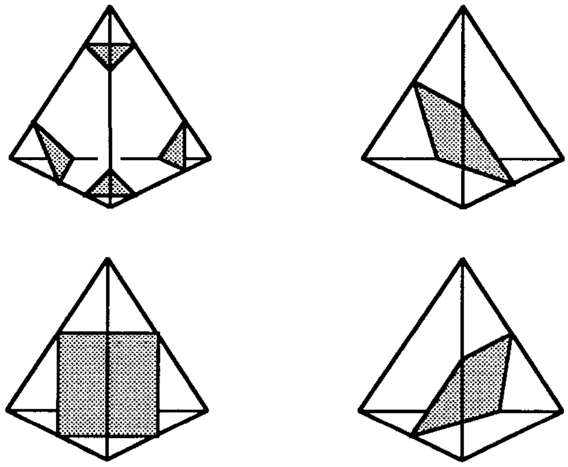


Fig. 11.

for H_1. The 3-simplices of $\mathcal{T}$ are the balls that are produced when H_1 is cut up along Δ_1. The graph Γ appears on the boundary of each 3-simplex τ as the dual (tetrahedral) graph to the 1-skeleton τ_1 of the tetrahedron τ.

Consider how a normal surface intersects the boundary of τ. A single component c of $F \cap \tau_2$ can run only once along any edge of Γ, or, put another way, c can cross an edge of τ_1 at most once. In particular c meets each face of τ in a single spanning arc (i.e., an arc whose ends lie on different sides of the triangular face). It follows immediately that a tetrahedron has up to normal isotopy precisely seven curve types. (See Figure 11.)There are four curve types with three sides and three curve types with four sides.

If α is a curve type in τ, and p is a point in the interior of τ, then the cone p*α of α to p is called a *disk type* of τ. Hence a tetrahedron has up to normal isotopy precisely seven disk types. We conclude that $F \subset M$ is a normal surface if and only if F intersects each tetrahedron of $\mathcal{T}$ in a (necessarily pairwise disjoint) collection of these disktypes.

Thus a normal surface is determined by the number of each curve type in which it meets the boundaries of the various tetrahedra. That is, if $\mathcal{C}_1, \ldots, \mathcal{C}_n$ is an ordering of the curve types, then the surface F determines (and is determined by) an n-tuple $(x_1, \ldots, x_n)$, where x_i denotes the number of representatives of $\mathcal{C}_i$ which F induces in the tetrahedra of $\mathcal{T}$.

Conversely, if we start with an n-tuple of non-negative integers, then we can construct a normal surface in M corresponding to this n-tuple if it satisfies the following constraints:

(i) We cannot have two 4-sided disks from distinct normal isotopy classes in the same tetrahedron (for they necessarily intersect).

(ii) Edges of disktypes on corresponding faces of incident tetrahedra have to match. Namely, if F intersects one face of a tetrahedron in p representatives of a certain arc type, then F also has to intersect the corresponding face of the incident tetrahedron in p representatives of the same arc type.

We now explain "geometric" addition of normal surfaces. A normal surface F in M is *straight* if it satisfies these conditions:

(i) For any 2-simplex σ in T^2, $\sigma \cap F$ consists only of straight spanning arcs (called *chords*).
(ii) In each tetrahedron τ any 3-sided disk in $\tau \cap F$ is the triangle given by the convex hull of its vertices.
(iii) Any 4-sided disk in $\tau \cap F$ is the cone to the barycenter of its four vertices.

Clearly any normal surface can be isotoped to be straight. Now consider how two straight normal surfaces F_1 and F_2 intersect. First move them slightly so that $F_1 \cap F_2 \cap T^1 = \emptyset$ and so that no barycenter of a 4-sided disk in F_2 lies in F_1 (and vice versa). Then

LEMMA 8.4. *In each tetrahedron τ, $F_1 \cap F_2$ consists of proper arcs, each of which has its ends on distinct 2-simplices. Each end is a point in a 2-simplex $\sigma \prec \tau$ where a chord of $F_1 \cap \sigma$ and a chord of $F_2 \cap \sigma$ intersect.*

Consider how chords in a 2-simplex σ can intersect. Let p be the intersection point. There is a unique way to remove an X neighborhood of p and rejoin the endpoints of the X by two disjoint arcs so that the result gives two spanning arcs in σ. This process is called a *regular exchange* at p.

Now consider extending this regular exchange along an arc component C of $F_1 \cap F_2$ inside a tetrahedron. That is, given two straight disks in a tetrahedron which intersect along an arc C, try to remove a neighborhood of C from both F_1 and F_2 and reattach the sides so that the result is a regular exchange at the ends of C. It is easy to see that this is possible, *unless* the disk types are distinct and both 4-sided.

We say that normal surfaces F_1 and F_2 are *compatible* if, in each tetrahedron, the four-sided curve types of F_1 and F_2 (if any) are the same. If F_1 and F_2 are compatible then, after they are straightened, we have seen that in a neighborhood of each curve of $F_1 \cap F_2$ it is possible to perform a regular exchange to eliminate the curve of intersection. The result of this operation on all intersection curves is a normal surface called the *geometric sum* of F_1 and F_2. Denote this surface by $F_1 + F_2$.

There are several interesting properties which are additive with respect to the geometric sum operation.

If F_1 and F_2 are compatible normal surfaces, then $F_1 + F_2$ is defined and
(i) $\chi(F_1 + F_2) = \chi(F_1) + \chi(F_2)$, where χ is Euler characteristic.
(ii) If F_1 corresponds to $(x_1, \dots, x_n)$ and F_2 corresponds to $(y_1, \dots, y_n)$, then $F_1 + F_2$ corresponds to $(x_1 + y_1, \dots, x_n + y_n)$.
(iii) $w(F_1 + F_2) = w(F_1) + w(F_2)$, where $w(F) =$ weight of $F = |F \cap T^1|$.

Now it is easy to see that the solution set of a system of integral equations in the positive orthant is generated under addition by a finite number of "fundamental" solutions which can be found algorithmically (see, e.g., [19, Chapter 8]). Exploiting the properties of the geometric sum listed above, it's often possible to show that if any surface with a compression preserved property appears in M then one with this property appears among the fundamental surfaces. If it can then be checked whether each of the fundamental surfaces has the desired property, the result is an algorithm to decide if M contains a surface with the desired property. So, for example, there is an algorithm to detect the presence of a reducing sphere, and an algorithm to detect the presence of an injective surface (see [22,2]). Part of this problem requires recognizing if a 2-sphere is a reducing sphere or whether it

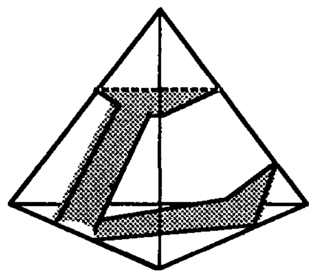


Fig. 12.

bounds a 3-ball (see [51]). This is a difficult problem in its own right, and one that requires a new idea – that of an "almost normal surface". Such a surface is normal, except in a single tetrahedron whose boundary it intersects in an octagon. (See Figure 12.) This leads us into an area of very active research. For example, see [50] for a discussion of how strongly irreducible splitting surfaces can be put in almost normal position and see [37] for a provocative discussion of other algorithms which may be useful and which make use of almost normal surfaces.

References

[1] S. Akbulut and J. McCarthy, *Casson's Invariant for Oriented Homology 3-Spheres: An Exposition*, Mathematical Notes, Vol. 36, Princeton University Press (1990).

[2] A. Bart and M. Scharlemann, *Least weight injective surfaces are fundamental*, Topology Appl. **69** (1996), 251–264.

[3] R.H. Bing, *An alternative proof that 3-manifolds can be triangulated*, Ann. of Math. **69** (1959), 37–65.

[4] J. Birman, *On the equivalences of Heegaard splittings of closed orientable 3-manifolds*, Knots, Groups and 3-Manifolds, L.P. Neuwirth, ed., Ann. Math. Studies, Vol. 84, Princeton University Press (1975).

[5] J. Birman, F. González-Acuña and J. Montesinos, *Heegaard splittings of prime 3-manifolds are not unique*, Mich. Math J. **23** (1976), 97–103.

[6] M. Boileau and J.P. Otal, *Groupes des difféotopies de certaines varietés de Seifert*, C. R. Acad. Sci. Paris **303-I** (1986), 19–22.

[7] M. Boileau and J.P. Otal, *Sur les scindements de Heegaard du tore* T^3, J. Differential Geom. **32** (1990), 209–233.

[8] M. Boileau and H. Zieschang, *Heegaard genus of closed orientable Seifert 3-manifolds*, Invent. Math. **76** (1984), 455–468.

[9] F. Bonahon, *Difféotopies des espaces lenticulaires*, Topology **22** (1983), 305–314.

[10] F. Bonahon and J.P. Otal, *Scindements de Heegaard des espaces lenticulaires*, Ann. Scient. Éc. Norm. Sup. **16** (1983), 451–466.

[11] S. Boyer, *Dehn surgery on knots*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 165–218.

[12] A. Casson and C.McA. Gordon, *Reducing Heegaard splittings*, Topology Appl. **27** (1987), 275–283.

[13] A. Casson and C.McA. Gordon, *Manifolds with irreducible Heegaard splittings of arbitrarily high genus* (unpublished).

[14] J. Cerf, *Sur les difféomorphismes de la sphère de dimension trois* ($\Gamma_4 = 0$), Lecture Notes in Math., Vol. 53, Springer, Berlin (1968).

[15] R. Engmann, *Nicht-homöomorphe Heegaard-Zerlegungen vom Geschlect 2 der zusammenhängenden Summer zweier Linsenräume*, Abh. Math. Sem. Univ. Hamburg **35** (1970), 33–38.

[16] W. Haken, *Theorie der Normalflächen*, Acta. Math. **105** (1961), 245–375.

[17] W. Haken, *Some results on surfaces in 3-manifolds*, Studies in Modern Topology, Math. Assoc. Am., Prentice Hall (1968), 34–98.

[18] P. Heegaard, *Forstudier til en topologisk teori for de algebraiske fladers sammenhaeng*, Ph.D. Thesis, Copenhagen (1898).

[19] J. Hemion, *The Classification of Knots and 3-Dimensional Spaces*, Oxford University Press (1992).

[20] J. Hempel, *3-Manifolds*, Annals of Math Studies, Vol. 86, Princeton University Press (1976).

[21] W. Jaco, *Lectures on Three-Manifold Topology*, Regional Conference Series in Mathematics, Vol. 43, Amer. Math. Soc., Providence, RI (1981).

[22] W. Jaco and U. Oertel, *An algorithm to decide if a 3-manifold is a Haken manifold*, Topology **23** (1984), 195–209.

[23] W. Jaco and J.H. Rubinstein, *PL minimal surfaces*, J. Differential Geom. **27** (1988), 493–524.

[24] K. Johannson, *Topology and Combinatorics of 3-Manifolds*, Lecture Notes in Math., Vol. 1599, Springer, Berlin (1995).

[25] T. Kobayashi, *A construction of 3-manifolds whose homeomorphism classes of Heegaard splittings have polynomial growth*, Osaka J. Math. **29** (1992), 653–674.

[26] M. Lustig and Y. Moriah, *Nielsen equivalence in Fuchsian groups and Seifert fibered spaces*, Topology **30** (1991), 191–204.

[27] W. Magnus, A. Karrass and D. Solitar, *Combinatorial Group Theory*, Dover (1966).

[28] J. Milnor, *Morse Theory*, Annals of Math Studies, Vol. 51, Princeton University Press (1963).

[29] E. Moise, *Affine structures in 3-manifolds, V: the triangulation theorem and Hauptvermutung*, Ann. of Math. **55** (1952), 96–114.

[30] J. Montesinos-Amilibia, *Discrepancy between the rank and the Heegaard genus of a 3-manifold*, Note Mat. 9 suppl. (1989), 101–117.

[31] Y. Moriah and J.H. Rubinstein, *Heegaard structures of negatively curved 3-manifolds*, Comm. Anal. Geom. **5** (1997), 375–412.

[32] Y. Moriah and J. Schultens, *Irreducible Heegaard splittings of Seifert fibered spaces are either vertical or horizontal*, Topology **37** (1998), 1089–1112.

[33] J.-P. Otal, *Sur les scindements de Heegaard de la sphere* S^3, Topology **30** (1991), 249–258.

[34] Parris, *Pretzel knots*, Ph. D. Thesis, Princeton University (1978).

[35] A. Przybyszewska, in *Knot theory from Vandermonde to Jones* by J. Przytycki, Preprint 43, Mathematics Institute, Odense University (1993).

[36] C. Rourke and B. Sanderson, *Piecewise Linear Topology*, Springer, Berlin (1982).

[37] J.H. Rubinstein, *Polyhedral minimal surfaces and decision problems for 3-manifolds*, Regional Conference Series in Mathematics, Amer. Math. Soc. (1996), to appear.

[38] J.H. Rubinstein and M. Scharlemann, *Comparing Heegaard splittings of non-Haken 3-manifolds*, Topology **35** (1996), 1005–1026.

[39] J.H. Rubinstein and M. Scharlemann, *Transverse Heegaard splittings*, Michigan Math. J. **44** (1997), 69–83.

[40] J.H. Rubinstein and M. Scharlemann, *Comparing Heegaard splittings of non-Haken 3-manifolds – the bounded case*, Trans. Amer. Math. Soc. **350** (1998), 689–715.

[41] M. Scharlemann and A. Thompson, *Heegaard splittings of (surface)* $\times$ *I are standard*, Math. Ann. **295** (1993), 549–564.

[42] M. Scharlemann and A. Thompson, *Thin position and Heegaard splittings of the 3-sphere*, J. Differential Geom. **39** (1994), 343–357.

[43] M. Scharlemann and A. Thompson, *Thin position for 3-manifolds*, AMS Contemporary Math., Vol. 164 (1994), 231–238.

[44] J. Schultens, *The classification of Heegaard splittings for (compact orientable surfaces)* $\times S^1$, Proc. London Math. Soc. **67** (1993), 425–448.

[45] J. Schultens, *Heegaard splittings of Seifert fibered spaces with boundary*, Trans. Amer. Math. Soc. **347** (1995), 2533–2552.

[46] J. Schultens, *The stabilization problem for Heegaard splittings of Seifert fibered spaces*, Topology Appl. **73** (1996), 133–139.

[47] P. Scott, *The eight geometries of 3-manifolds*, Bull. Lond. Math. Soc. **15** (1983), 401–487.

[48] E. Sedgwick, *An infinite collection of Heegaard splittings that are equivalent after one stabilization*, Math. Ann. **308** (1997), 65–72.

[49] E. Sedgwick, *The irreducibility of Heegaard splittings of Seifert fibered spaces*, to appear.

[50] M. Stocking, *Almost normal surfaces in* 3*-manifolds*, Preprint.

[51] A. Thompson, *Thin position and the recognition problem for* S^3, Math. Res. Lett. **1** (1994), 613–630.

[52] F. Waldhausen, *Heegaard-Zerlegungen der* 3*-Sphäre*, Topology **7** (1968), 195–203.

[53] H. Zieschang, *On Heegaard diagrams of* 3*-manifolds*, Asterisque **163–164** (1988), 247–280.

CHAPTER 19

Representations of 3-Manifold Groups

Peter B. Shalen*

Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Contents

*Supported by an NSF grant and a U. of I. University Scholar award.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

The phrase "representations of 3-manifold groups" is likely to suggest different things to different people. When I was asked to write a chapter for this handbook on this subject, I thought it would be wise to focus on something I knew about, and I therefore decided to concentrate on the interaction between two kinds of representations of fundamental groups of 3-manifolds: representations by 2×2 matrices, and representations by automorphisms of trees. Representations of the first kind are related to hyperbolic structures on 3-manifolds, while those of the second kind are related to surfaces in 3-manifolds. The interaction between these two kinds of representations therefore provides a link between what are probably the two most useful kinds of structures on 3-manifolds, whence the utility of this theory in studying group actions, Dehn surgery, surfaces in knot exteriors, and degeneration of hyperbolic structures.

The core of this chapter is an attempt to present – with the necessary background – some of the content of a series of joint papers [20–22] by Marc Culler and myself, and of the first chapter of our paper [18] with Cameron Gordon and John Luecke. I have been able to give only the briefest hints about the remarkable developments in the area that have been made by Steven Boyer and Xingru Zhang in [8–10], by Daryl Cooper and Darren Long and their co-authors in [14,16], and other papers – see [17] for a survey – and by Nathan Dunfield in [25] and [26]. I have been even briefer about my closely related joint work with John Morgan in [45–47], and the subsequent breakthrough by Rips and the developments in geometry group theory that it has led to. To do justice to this material would have made the chapter twice as long, but of course the point of this enterprise is to try to inspire you to read all these papers. By and large the material that I have discussed in detail is prerequisite for the material that I have touched on lightly, and naturally it's material that I know well.

The chapter is meant to be organized rather as if it were the notes for a course, and I've tried to keep the informal tone of a course or lecture series, as I did in my survey articles [60] and [61]. As the main topic of the chapter involves tying together ideas from several areas of mathematics other than topology, I have tried to provide reasonable introductions to the relevant ideas from these other areas. In some case this has meant not just introducing statements or even just proofs of relevant theorems, but trying to provide some real context by showing how these ideas are used within the areas from which they're borrowed. That's why you'll find little sections about topics like Ihara's theorem on discrete subgroups of $\mathrm{SL}(2, \mathbf{Q}_p)$ and Lagrange's theorem that every positive integer is a sum of four squares.

The different sections are meant to be read in order, but in case you don't like being kept in suspense, I'll give a very quick outline of what's going to happen. (Of course I'll have to use some terms that may not mean much until you've read the relevant sections, but sometimes just seeing key words can be of value.) In Section 1 I'll present some generalities about group actions, particularly actions of fundamental groups, and I show how surfaces and hyperbolic structures lead naturally to actions on trees and representations in $\mathrm{SL}(2, \mathbf{C})$. In Section 2 I'll present a construction that goes the opposite way – starting with an action of π_1 of a 3-manifold on a tree, you can get a surface in the manifold. This idea is basically due to Stallings, and has lots of direct applications to 3-manifold theory, of which I'll do some samples.

Section 3 will be basically a mini-course on the tree for SL_2, which is a special case of the Bruhat–Tits building of an algebraic group that was given an elegant self-contained treatment by Serre. Like Serre, in his longer course [58], I'll assume only the most elementary algebra. This section provides the germ of the connection between matrix representations and representations by tree automorphisms. Before Culler and I wrote [20], this material had already been applied to 3-manifold theory via Bass's GL_2 subgroup theorem [3]. The applications in [20] and the papers depending on it differ from this first application in that they involve considering entire families of representations, which form algebraic varieties. In Section 4 I'll do a little rudimentary algebraic geometry, referring you to standard texts for some of the harder results, and I'll introduce varieties of representations of groups, and the varieties of characters that are closely related to them. I'll say a bit about what these things look like for the case of fundamental groups of hyperbolic 3-manifolds.

In Section 5 I'll present the basic theory that Culler and I developed in [20], tying together the material from Sections 3 and 4. In the last subsection of Section 5, and in Section 6, I'll do a first application to topology – the existence of an essential separating surface in the complement of a nontrivial knot, first conjectured by Neuwirth. A second application, in Section 7, is a proof of the Smith Conjecture about periodic tame homeomorphisms of S^3 with 1-dimensional fixed point set. (I'll talk a bit about the history of this when the time comes.) In Section 8 I'll introduce some machinery that's needed for the applications to Dehn surgery that I'll talk about in Section 9, and for studying more refined questions related to the Neuwirth Conjecture in Section 10. In Section 11 I'll give a hint about how the techniques are related to geometric questions about degenerations of hyperbolic structures.

I'm very grateful to Marc Culler for helping talk me through some difficult spots in the writing. (Of course, if Marc hadn't done all this joint work with me I wouldn't have anything to write about.) I'm very grateful to Benson Farb and Bob Daverman for reading the entire manuscript and making a huge number of helpful comments. I'd also like to thank Jeremy Teitelbaum for trying to make sure I didn't say anything too silly in the passages about p-adic numbers.

Although this chapter is something very different from what I had in mind when I sat down to write it, I hope it may prove useful.

1. Some basic concepts and examples

1.1. *Representations and actions*

Recall that an *action* of a group Γ on a set X is a function $\cdot : \Gamma \times X \to X$ that satisfies the identities $1 \cdot x = x$ and $(\gamma\delta) \cdot x = \gamma \cdot (\delta \cdot x)$. There is a lot of structure associated in an elementary way with an action, such as the partition of X into *orbits*: two elements x and y are in the same orbit if and only if $\gamma \cdot x = y$ for some $\gamma \in \Gamma$. The *stabilizer* of an element $x \in X$, often denoted Γ_x, is the subgroup of Γ consisting of all $\gamma \in \Gamma$ such that $\gamma \cdot x = x$. For any $x \in X$, the map $\Gamma \to X$ defined by $\gamma \to \gamma \cdot x$ induces a bijection between the set of cosets of the form $\gamma\Gamma_x$, for $\gamma \in \Gamma$, and the orbit $\Gamma \cdot x$ of x. A set $S \subset X$ is *invariant* under the action if it is a union of orbits, i.e., if $\gamma \cdot x \in S$ whenever $x \in S$ and $\gamma \in \Gamma$. An element of X is said to be *fixed* by Γ if $\{x\}$ is invariant, i.e., is an entire orbit. The action is *free* if the stabilizer of every element of X is the trivial subgroup of Γ.

There's a natural bijective correspondence between actions of Γ on X and *representations* of Γ in the symmetric group $\mathcal{S}(X)$, i.e., homomorphisms $\rho : \Gamma \to \mathcal{S}(X)$. If $\cdot$ is an action, the corresponding representation sends γ to the element $s \mapsto \gamma \cdot s$ of $\mathcal{S}(X)$. I'll generally be talking about sets X that have some extra structure, and focusing on actions which preserve this structure in the sense that the corresponding representations take values in the automorphism group $\mathrm{Aut}(X)$. Thus X may be a vector space over some field K, in which case $\mathrm{Aut}(X)$ is the group of linear automorphisms; in terms of the action this means that the identities $\gamma \cdot (x + y) = \gamma \cdot x + \gamma \cdot y$ and $\gamma \cdot (ax) = a\gamma \cdot x$ hold when $x, y \in X$ and $a \in K$. So we can talk about linear actions and linear representations, and the difference between the two is purely notational. Similarly we can talk about topological actions, or actions by homeomorphisms, of Γ on a topological space X; simplicial actions on a simplicial complex; and so forth.

A representation $\Gamma \to \mathrm{Aut}(X)$ is termed *faithful* if it is an injective homomorphism. An action is said to be *effective* if it corresponds to a faithful representation. Thus Γ acts effectively on X if and only if for every $\gamma \in \Gamma - \{1\}$ there is an $x \in X$ with $\gamma \cdot x \neq x$.

Suppose that we are given actions of a group Γ on two sets X and Y. A map of sets $f : X \to Y$ is said to be *Γ-equivariant* if we have $f(\gamma \cdot x) = \gamma \cdot f(y)$ for every $x \in X$. For any group Γ, there is a category in which the objects are sets equipped with actions of Γ and the morphisms are Γ-equivariant maps. By giving the sets, actions and maps extra structure, one can define many natural subcategories. For example, one can require the sets to be vector spaces, or topological spaces, or simplicial complexes, and require both actions and maps to be linear, or continuous, or simplicial. There are natural terms to designate isomorphisms in these categories: Γ-equivariant linear isomorphisms, Γ-equivariant homeomorphisms, Γ-equivariant simplicial isomorphisms and so on.

Of course, we think of two actions as being "the same" if we have an isomorphism (in the appropriate category) between the sets in question which is equivariant with respect to the given actions. I'll express this by saying that the actions – or the corresponding representations – are *equivalent*. (This is the classical term in the case of linear representations.) In more direct terms, $\rho : \Gamma \to \mathrm{Aut}(X)$ and $\rho' : \Gamma \to \mathrm{Aut}(Y)$ are equivalent if and only if there is an isomorphism $\phi : X \to Y$ such that $\phi \circ \rho = \rho' \circ \phi$.

By an *invariant* of a representation one means any sort of datum associated with the representation which depends only on its equivalence class. I'll illustrate the idea by talking briefly about invariants of linear representations, which, besides being a very classical thing to look at, will be especially important in this chapter.

The most obvious invariant of a linear representation $\rho : \Gamma \to \mathrm{Aut}(V)$, where V is a vector space over some given field, is the dimension of V, which is often called the dimension of the representation. It's just about obvious that any n-dimensional representation is equivalent to a representation in $\mathrm{Aut}(\mathbf{C}^n) = \mathrm{GL}_n(\mathbf{C})$, and that two representations $\rho, \rho' : \Gamma \to \mathrm{GL}_n(\mathbf{C})$ are equivalent if and only if $\rho' = i_A \circ \rho$ for some matrix $A \in \mathrm{GL}_n(\mathbf{C})$; here I am denoting by i_A the inner automorphism $X \mapsto AXA^{-1}$ of $\mathrm{GL}_n(\mathbf{C})$. In particular, the property of being *unimodular*, i.e., of sending the entire group Γ into $\mathrm{SL}_n(\mathbf{C})$, depends only on the equivalence class of a representation $\rho : \Gamma \to \mathrm{GL}_n(\mathbf{C})$.

Let me now specialize to the case of unimodular representations of dimension 2; this is the case that will be important in this chapter, and, conveniently, the case that I am competent to discuss. By far the most important invariant of such a representation is its *character*.

The character of $\rho : \Gamma \to \mathrm{GL}_n(\mathbf{C})$ is the function $\chi : \Gamma \to \mathbf{C}$ defined by $\chi(\gamma) = \operatorname{trace} \rho(\gamma)$. It is very nearly true that the character is a *complete* invariant, which would mean that 2-dimensional unimodular representations with the same character were always equivalent. To see that this is not quite true, note, for example, any homomorphism of Γ into the group of matrices of the form $\left(\begin{smallmatrix} 1 & \lambda \\ 0 & 1 \end{smallmatrix}\right)$ has the same character as the *trivial representation* that sends the entire group to the identity matrix. It turns out that the only bad examples of this sort involve *reducible* representations. A representation of Γ in $\mathrm{SL}(2, \mathbf{C})$ is said to be reducible if it is equivalent to a representation by upper triangular matrices; in terms of the corresponding action of Γ on $\mathbf{C}^2$, this means that some 1-dimensional subspace is invariant under the action. You will find a proof of the following elementary result in [20]:

PROPOSITION 1.1.1. *Let Γ be any finitely generated group. If two unimodular representations ρ and ρ' of Γ in* $\mathrm{SL}(2, \mathbf{C})$ *have the same character, then either ρ and ρ' are equivalent or they are both reducible.*

There is presumably a similar result about $\mathrm{GL}_n(\mathbf{C})$, but one probably has to be more careful about the statement.

1.2. *A word about base points*

This chapter is about representations (or actions) of fundamental groups of connected 3-manifolds. (Incidentally, when I say "manifold" I always mean "manifold with (possibly empty) boundary". This seems to be the modern convention among topologists. A manifold is *closed* if it is compact and has empty boundary.)

In dealing with fundamental groups it is always necessary to decide what to do about base points. In many situations I will be talking about *equivalence classes* of representations (or actions) of the fundamental group of a connected manifold. The point I would like to make here is that in these situations, there is a strong sense in which the choice of a base point is irrelevant. To see why this is so, let's consider two points x and w in the connected manifold M. Any path α from x to w defines an isomorphism $I_\alpha : \pi_1(X, w) \to \pi_1(X, x)$: for any loop γ based at w we have $I_\alpha([\gamma]) = [\alpha * \gamma * \bar{\alpha}]$. If α and α' are two paths from x to w, the composition $I_{\alpha'} \circ I_\alpha^{-1}$ is the inner automorphism $i_{[\alpha' * \bar{\alpha}]}$ of $\pi_1(X, x)$. So although the isomorphism between $\pi_1(X, w)$ and $\pi_1(X, x)$ is not canonical, it is canonical modulo composition with inner automorphisms.

Now if we are given an action of $\pi_1(M, x)$ on a set X, and if $\rho : \pi_1(M, x) \to \mathcal{S}(X)$ is the corresponding representation, then for any path α from x to w we have a representation $\rho \circ I_\alpha : \pi_1(X, w) \to \mathcal{S}(X)$. If in place of α we consider another path α', we have $I_{\alpha'} = i_g I_\alpha$ for some $g \in \pi_1(M)$, and hence $\rho \circ I_{\alpha'} = i_{\rho(g)} \circ \rho \circ I_\alpha$. Since $\rho \circ I_{\alpha'}$ and $\rho \circ I_\alpha$ differ by post-composition with an inner automorphism, they are equivalent representations. So a representation of $\pi_1(M, x)$ defines an equivalence class of representations of $\pi_1(M, w)$. It's equally easy to see that this equivalence class depends only on the equivalence class of the given representation of $\pi_1(M, w)$, and that we get in this way an absolutely canonical bijection between equivalence classes of representations of $\pi_1(M, x)$ and equivalence classes of representations of $\pi_1(M, w)$. So one can talk without ambiguity about *equivalence classes of representations or actions of* $\pi_1(M)$, without specifying a base point.

There will be various other kinds of situations in this chapter where base points will be suppressed for a very similar reason. I will give a hint here and there to remind you of what is going on.

1.3. *The universal covering*

If M is a connected manifold of any dimension n, and x is a base point in M, the most basic example of an action of $\pi_1(M, x)$ is the usual action by homeomorphisms on the universal covering space $\widetilde{M}$. The uniqueness theorem for the universal covering tells us that this action is well-defined up to equivalence. (Here the underlying category is that of topological spaces. Thus two actions on spaces are equivalent if and only if there is an equivariant homeomorphism between the spaces.) Following the convention I've just explained, I will say – without mentioning a base point – that the action of $\pi_1(M)$ on $\widetilde{M}$ is canonically defined up to equivalence. If M is given a triangulation, then $\widetilde{M}$ inherits a triangulation, and we can then interpret the action of $\pi_1(M)$ as a simplicial action, defined up to equivalence in the simplicial category.

Giving other kinds of structure in the connected manifold M often leads to new actions of $\pi_1(M)$ that are induced by its action on $\widetilde{M}$.

1.4. *The tree associated with a hypersurface*

As one nice example, suppose that we are given a hypersurface in M, i.e., a codimension-1 submanifold F of M, not necessarily connected. In this section I will be assuming that F admits a *bicollaring,* i.e., a homeomorphism h of $F \times [-1, 1]$ onto a neighborhood of F in M such that $F(x, 0) = x$ for every $x \in F$ and $h(F \times [-1, 1]) \cap \partial M = h(\partial F \times [-1, 1])$. We can use a bicollaring h to define a partition of the space M into disjoint subsets. The subsets are of two types: the components of $M - h(F \times (-1, 1))$, and the sets of the form $F_i \times \{t\}$, where F_i is a component of F and $t \in (0, 1)$. We can regard the sets in this partition as forming a topological space with the quotient topology, and you will see easily that it is a graph, i.e., a 1-dimensional CW complex, which has one edge for every component of F and one vertex for every component of $M - F$. Note that h defines an identification of each edge of $\mathcal{G}$ with the interval $[-1, 1]$.

By construction there is a natural map $r : M \to \mathcal{G}$, but it is also very easy to construct a map $i : \mathcal{G} \to M$ such that the composition $r \circ i$ maps each edge and each vertex of $\mathcal{G}$ into itself. In particular $r \circ i$ is homotopic to the identity map of $\mathcal{G}$, from which it follows that $\mathcal{G}$ is connected and that $\pi_1(\mathcal{G})$ is isomorphic to a retract (hence both a subgroup and a quotient) of $\pi_1(M)$. The graph $\mathcal{G}$ is often called the *dual graph* of F in M. (Since regular neighborhood theory tells us that the bicollaring h is unique up to ambient isotopy, the graph is well-defined up to simplicial isomorphism, and even the map r is well-defined in a sense that's easy to work out.)

Now consider the universal covering $(\widetilde{M}, p)$ of M. Given the bicollaring h of F, it's a simple exercise in covering space theory to show that $\widetilde{F} = p^{-1}(F)$ has a unique bicollaring

$\tilde{h}$ in M such that $p(\tilde{h}(x,t)) = h(p(x),t)$ for all $x \in \widetilde{F}$, $t \in [0,1]$. Let T denote the dual graph of $\widetilde{F}$ in $\widetilde{M}$ defined in terms of this induced bicollaring $\tilde{h}$. Then T is simply connected since $\pi_1(T)$ is a retract of $\pi_1(\widetilde{M}) = \{1\}$; that is, T is a *tree*. Now the sets that make up the partition defining T are the components of the sets of the form $p^{-1}(A)$, where A ranges over the sets in the partition defining the dual graph $\mathcal{G}$. So the partition defining T is invariant under the action of $\pi_1(M)$, in the sense that each element of $\pi_1(M)$ maps each set in the partition onto a possibly different set in the partition. Hence the action of $\pi_1(M)$ on $\widetilde{M}$ induces an action on T. In fact, this induced action is the unique action that makes the quotient map $\widetilde{M} \to T$ equivariant.

Since T is the dual graph of $\mathcal{G}$ defined by the bicollaring $\tilde{h}$, each closed edge e of T comes equipped with an identification with $[-1,1]$, i.e., a homeomorphism $\eta_e : [-1,1] \to e$. Since T is simply connected, and therefore has no multiple edges, the identification of the edges of T with linear intervals give T the structure of a simplicial complex. Because of the precise way in which $\tilde{h}$ is induced from the bicollaring h of M, it's easy to check that for each element $\gamma \in \pi_1(M)$ and each edge e of T, and each $t \in [-1,1]$, we have $\gamma \cdot \eta_e(t) = \eta_{\gamma \cdot e}(t)$. This implies that the action of $\pi_1(M)$ on T is simplicial, but it also shows a little more – namely, that if an element $\gamma \in \pi_1(M)$ leaves an edge e of T invariant, then it actually fixes the edge pointwise. This property is expressed by saying that $\pi_1(M)$ acts on T *without inversions.* A simplicial automorphism of a tree is called an *inversion* if it leaves some edge invariant but interchanges its endpoints.

Because the bicollaring h is unique up to ambient isotopy, the tree T is well-defined up to simplicial equivalence once the hypersurface F is given.

It's easy to describe the stabilizers of the vertices and edges of T under the action of $\pi_1(M)$. Each vertex s of T corresponds to a component $\widetilde{K}$ of $\widetilde{M} - \tilde{h}((-1,1))$, and it follows from the construction that the stabilizer $\pi_1(M)_s$ of s coincides with the stabilizer $\pi_1(M)_{\widetilde{K}}$ of $\widetilde{K}$. But $\widetilde{K}$ is a component of $p^{-1}(K_j)$ for some component K_j of $M - h((-1,1))$, and from covering space theory one knows that the stabilizers of the various components of $p^{-1}(K_j)$ are precisely the conjugates of $\mathrm{im}(\pi_1(K_j)) \to \pi_1(M)$ in $\pi_1(M)$. (It follows from the kind of thing I talked about in Section 1.2 that the subgroup $\mathrm{im}(\pi_1(K_j)) \to \pi_1(M)$ of $\pi_1(M)$ is well-defined up to conjugacy, so it makes sense to talk about "conjugates of $\mathrm{im}(\pi_1(K_j)) \to \pi_1(M)$ in $\pi_1(M)$".) Furthermore, we have $\mathrm{im}(\pi_1(K_j)) \to \pi_1(M) = \mathrm{im}(\pi_1(C_j)) \to \pi_1(M)$, where C_j is the component of $M - F$ containing K_j. So we see that the stabilizers of the vertices of T are precisely the conjugates of the subgroups $\mathrm{im}(\pi_1(C_j)) \to \pi_1(M)$ in $\pi_1(M)$, where C_j ranges over the components of $M - F$. Similarly, the stabilizers of the edges of T are precisely the conjugates of the subgroups $\mathrm{im}(\pi_1(F_i)) \to \pi_1(M)$ in $\pi_1(M)$, where F_i ranges over the components of F.

The picture of the action is especially nice in the case where the inclusion homomorphism $\pi_1(F_i) \to \pi_1(M)$ is injective for every component F_i of F. For one thing, the stabilizer of an edge of T is then actually isomorphic to F_i, the isomorphism being canonical up to composition with an inner automorphism. We have a similarly nice situation with regard to the vertex stabilizers. This is because the injectivity of the inclusion homomorphisms $\pi_1(F_i) \to \pi_1(M)$ implies that for every component C_j of $M - F$, the inclusion homomorphism $\pi_1(C_j) \to \pi_1(M)$ is injective. You can see this very neatly in terms of the above discussion of $\widetilde{M}$: the injectivity of the homomorphisms $\pi_1(C_j) \to \pi_1(M)$ is equivalent to the assertion that the components of $\widetilde{M} - \widetilde{F}$ are simply connected. But the injectivity of the $\pi_1(F_i) \to \pi_1(M)$ implies that the components of $\widetilde{F}$ are simply connected, and since $\widetilde{M}$ is simply connected as well, the simple connectivity of the components of $\widetilde{M} - \widetilde{F}$ follows from van Kampen's theorem. Now, from the injectivity of the $\pi_1(C_j) \to \pi_1(M)$ and the general discussion above, we see that the stabilizers in $\pi_1(M)$ of the vertices of T are isomorphic to the groups $\pi_1(C_j)$ for components C_j of $M - F$. Again the isomorphisms are canonical up to composition with inner automorphisms.

Having shown how a bicollared hypersurface gives rise to an action on a tree, I should say a word about why the condition that a hypersurface be bicollared is a natural one. The obvious necessary conditions for a hypersurface F of M to be bicollared are that F be *properly embedded* and *two-sided*. To say that F is properly embedded in M means that F is a closed subset of M and that $F \cap \partial M = \partial F$. To say that a properly embedded hypersurface F is two-sided (or "locally separating") means that the complement of F relative to any sufficiently small neighborhood of F is disconnected. (For simplicity, or out of habit, I will be talking mostly about orientable manifolds in this chapter. If M is orientable then a properly embedded hypersurface $F \subset M$ is two-sided if and only if it is orientable – which means, if you like, that each component is orientable.)

Conversely, any halfway-reasonable properly embedded orientable hypersurface in M is bicollared: for example, if M comes with a smooth or piecewise-linear structure then any smooth or PL hypersurface in M which is orientable or properly embedded is bicollared. (If you don't make some assumption about the surface then you can run into weird examples like the Alexander horned sphere [42].) For technical reasons it is not always convenient to be working with a smooth or PL structure, so I'll often just make it a hypothesis that the surfaces I talk about are bicollared.

1.5. *The three-dimensional case: Essential surfaces*

The construction I described in the last subsection is especially useful in the case $n = 3$. Before making a few simple comments about this case I need to make some basic definitions and remarks about surfaces in 3-manifolds which are important for everything I'll be talking about in this chapter.

A 3-manifold M is said to be *irreducible* if M is connected and every bicollared 2-sphere in M is the boundary of a 3-ball contained in M. A bicollared surface $F \subset M$ is said to be *boundary-parallel* if F is the frontier of a set $P \subset M$ such that the pair (P, F) is homeomorphic to $(F \times [0, 1], F \times \{1\})$. (The *frontier* of a set is the intersection of its closure with the closure of its complement; I am reserving the term *boundary* for the intrinsic sense, as in "manifold with boundary". Note that in the definition I've just given, the homeomorphism of P onto $F \times [0, 1]$ has to map $P \cap \partial M$ onto $(\partial F \times [0, 1] \cup (F \times \{0\}).$)

DEFINITION 1.5.1. For most of this chapter I'll be using the following definition. A surface F in a compact, irreducible, orientable 3-manifold is said to be *essential* if it has the following properties:

(i) F is bicollared;
(ii) the inclusion homomorphism $\pi_1(F_i) \to \pi_1(M)$ is injective for every component F_i of F;
(iii) no component of F is a 2-sphere;
(iv) no component of F is boundary-parallel; and
(v) F is nonempty.

Condition 1.5.1(ii) has a beautiful geometric interpretation. One situation in which the condition obviously *fails* to hold is the one in which there is a disk $D \subset M$ such that (a) $D \cap F = \partial D$, and (b) the simple closed ∂D is homotopically nontrivial in F – which by elementary surface topology (see [29, Theorem I.7]) is the same as saying that it doesn't bound a disk in F. In this case, ∂D clearly defines (up to conjugation and inversion) a nontrivial element of $\ker(\pi_1(F_i) \to \pi_1(M))$, where F_i is the component of F containing ∂D. Now, conversely, it is a fundamental principle in 3-manifold theory [34, proof of Lemma 6.1], which is easily deduced from two results due to Papakyriakopoulos, the Dehn Lemma and the Loop Theorem, that if a bicollared surface F fails to satisfy condition 1.5.1(ii), then there is a disk D satisfying (a) and (b).

Property (a) of the disk D says it can be thought of as a properly embedded disk in the manifold M' obtained by splitting M along F. The proof of Papakyriakopoulos's theorem is usually done in the PL category, and gives the additional conclusion that (c) D is bicollared in M'. A disk satisfying (a), (b) and (c) is often called a *compressing disk* for F. Thus, for a bicollared surface F, condition 1.5.1(ii) is equivalent to the condition that F there is no compressing disk for F.

The properties that I have included in the definition of an "essential surface" are very similar to those that people typically include in the definition of an "incompressible surface". However, Haken's term "incompressible" has been used in so many ways in recent years that there are now almost as many definitions as there are 3-manifold topologists, and – to make matters worse – people get emotional about the issue of what the term should mean. That's why I am avoiding it in this chapter.

According to the discussion at the end of Section 1.4, condition 1.5.1(i) implies that the stabilizer of each edge of T is isomorphic to the fundamental group of some component F_i of F, and that the stabilizer in $\pi_1(M)$ of each vertex of T is isomorphic to the fundamental group of some components C_j of $M - F$; and that these isomorphisms are canonical up to composition with inner automorphisms.

Conditions 1.5.1(ii)–(iv) in the definition of an essential surface also give nice information about the action associated to the surface. A (simplicial) action (without inversions) of a group Γ on a tree T is said to be *trivial* if there is a vertex of T which is fixed by the entire group Γ.

PROPOSITION 1.5.2. *Let F be an essential surface in a compact, connected, orientable, irreducible 3-manifold M. Then the action on a tree associated to F is nontrivial.*

PROOF. I'll call the tree T. Assume that the action is trivial, so that the stabilizer of some vertex of T is all of $\pi_1(M)$. This translates into saying that for some component C of $M - F$, the injection $\pi_1(C) \to \pi_1(M)$ is an isomorphism (as in "isomorphism onto").

By condition 1.5.1(iv) we have $F \neq \emptyset$. Let F_0 be a component of F. Since $F_0 \cap C = \emptyset$, there is some component C_0 of $M - F_0$ such that the inclusion homomorphism $\pi_1(C_0) \to \pi_1(M)$ is surjective. This gives a contradiction right off the bat if F_0 doesn't separate M, i.e., if C_0 is the only component of $M - F_0$, because then even $H_1(M - F_0) \to H_1(M)$ fails to be surjective. Suppose now that $M - F_0$ has a second component C_1. Consider the dual tree T' to the connected essential surface F_0. If e is any edge of T', one endpoint s_0 of e, corresponding to a component of $p^{-1}(C_0)$, is stabilized by all of $\pi_1(M)$. The other endpoint s_1 of e has stabilizer $\pi_1(M)_{s_1} = \pi_1(M)_{s_1} \cap \pi_1(M)_{s_0} = \pi_1(M)_e$. This means that the inclusion homomorphism $\pi_1(F) \to \pi_1(\overline{C_1})$ is an isomorphism.

We also know that the $\overline{C_i}$ are irreducible; otherwise, since M is irreducible, F_0 would be contained in a ball, and would not be essential. Now the main point is to apply a theorem due to Stallings, for which the best reference is Brown and Crowell's paper [11] in which a stronger theorem is proved: Stallings's theorem says that if K is a compact, connected, orientable, irreducible 3-manifold and $F \subset K$ is a connected 2-manifold such that $\pi_1(F) \to \pi_1(K)$ is an isomorphism, then either K is a ball and $F = \partial K$, or the pair (K, F) is homeomorphic to $(F \times [0, 1], F \times \{1\})$. We can apply this with F_0 and $\overline{C_1}$ in place of F and K, and we get a contradiction to either 1.5.1(ii) or 1.5.1(iii). □

So an essential surface in M gives rise to a nontrivial action (defined up to equivalence) of $\pi_1(M)$ on a tree. In Section 2 we'll see how to go the other way – to start with an action of $\pi_1(M)$ on a tree and use it to construct, in a noncanonical but ultimately very useful way, an essential surface in M.

1.6. *The action associated to a hyperbolic structure*

Another kind of structure in a connected manifold M (of dimension $n \geqslant 2$) that leads to an action of the fundamental group is a (complete) hyperbolic structure. I will refer you to Bonahon's chapter in this volume for an introduction to hyperbolic geometry, which plays an important role in most of the topics I will be covering in this chapter. If we are given a hyperbolic structure on M, we can identify the universal covering of $\widetilde{M}$ with the hyperbolic space $\mathbf{H}^n$ by some isometry. The natural action of $\pi_1(M)$ then becomes an action by isometries on $\mathbf{H}^n$ – or, in slightly different language, a representation ρ_0 of $\pi_1(M)$ in the isometry group Isom($\mathbf{H}^n$). The representation that we get in this way is well-defined up to equivalence once we have specified a hyperbolic structure on M. The representation is readily seen to be faithful, and to be *discrete* in the sense that $\rho_0(\pi_1(M))$ is a discrete subgroup of the topological group Isom($\mathbf{H}^n$). If M is orientable then ρ_0 takes values in the group Isom$^+$($\mathbf{H}^n$) of orientation-preserving isometries, but it is still well-defined only up to conjugation in Isom($\mathbf{H}^n$). Thus in general there may be two inequivalent representations in Isom$^+$($\mathbf{H}^n$) associated to a given hyperbolic structure on M, and they will differ by conjugation by an orientation-reversing involution $J \in$ Isom$^+$($\mathbf{H}^n$), a reflection about a hyperplane.

In the three-dimensional case, Isom($\mathbf{H}^3$) may be identified isomorphically with PSL(2, **C**). The identification is canonical modulo inner automorphisms. An element of PSL(2, **C**) is a coset modulo $\pm I$ of a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. I'll denote this coset by $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. We may choose the orientation-reversing involution J so that the group-theoretical conjugation $A \to A^J$ is realized by complex conjugation of matrices:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \longmapsto \begin{bmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{bmatrix}.$$

Thus the two equivalence classes of representations in $\mathrm{PSL}(2, \mathbf{C})$ associated with a given hyperbolic structure differ from each other by a complex conjugation.

Sometimes it's more convenient to work with representations in $\mathrm{SL}(2, \mathbf{C})$ instead of $\mathrm{PSL}(2, \mathbf{C})$. For this purpose, it is useful to have the following result of Thurston's:

PROPOSITION 1.6.1. *Let M be a (connected) orientable hyperbolic 3-manifold, and let $\rho_0 : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ be a representation associated to the hyperbolic structure. (So ρ_0 belongs to one of the two equivalence classes discussed above.) Then there is a lift of ρ_0 to $\mathrm{SL}(2, \mathbf{C})$, i.e., a representation $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ such that $p\tilde{\rho}_0 = \rho_0$, where $p : \mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is the quotient projection.*

The representation $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ is even less canonical than the representation ρ_0. Whereas there are in general two choices for ρ_0 in terms of a given hyperbolic structure on M, it is a simple exercise, given Proposition 1.6.1, to show that when $\pi_1(M)$ is finitely generated, the number of lifts of a given ρ_0 is $|H_1(M; \mathbf{Z}_2)|$. However, a lift of ρ_0, being a linear representation, is a pretty down-to-earth kind of object, and one from which the hyperbolic structure itself can be recovered. For these reasons, having Proposition 1.6.1 is often very convenient in applications, as we shall see.

2. Actions of 3-manifold groups on trees

The ideas in this section are mostly due to Stallings, who developed them in a series of papers beginning with [64], and presented some of them in his book [65]. I'll be presenting them from a point of view which is fairly close to the one that was used in [47] and, a little later, in [18]. This point of view is influenced by Serre's book [58].

In this section, M will denote a compact, orientable, irreducible 3-manifold. In Section 1.4 I described how an essential surface in M gives rise to a nontrivial (simplicial) action, without inversions, of $\pi_1(M)$ on a tree. The construction of the action from the surface is canonical up to equivalence. It is far from being true that every nontrivial action without inversions of $\pi_1(M)$ on a tree arises from this construction. Indeed, I pointed out that under the action associated with an essential surface F, the stabilizer of each edge or vertex of the tree is isomorphic to the fundamental group of a component of F or of $M - F$. By contrast, I will point out in Section 2.3 below that for many reasonable choices of M there are very simple and natural examples of nontrivial actions without inversions of $\pi_1(M)$ on trees for which the edge and vertex stabilizers are not even finitely generated!

Nevertheless, it turns out that with every nontrivial action without inversions of $\pi_1(M)$ on a tree one can "associate", in an interesting way, an essential surface in M. I've used quotation marks here because the construction of the surface from the action is far from being canonical, as it depends on many choices. Furthermore, one cannot in general reconstruct an action from a surface "associated" to it. Nevertheless, such a surface contains important information about the action. In this section I will describe the construction of a surface associated to an action, talk a little about the extent to which this construction behaves like an inverse to the construction of 1.4, and give some important applications. These are only first applications, though, because everything I will be talking about in the rest of the chapter depends on the construction I'll describe here.

2.1. *Constructing an equivariant map*

Suppose, then, that we're given a simplicial action $\cdot$ of $\pi_1(M)$ on a tree T. (Eventually it will matter that the action is nontrivial and without inversions; I'll point out where these hypotheses come up.) The first step in constructing an essential surface associated with the action is to construct a (continuous) $\pi_1(M)$-equivariant map $\tilde{f} : \widetilde{M} \to T$.

Let's fix a triangulation for M, and give $\widetilde{M}$ the induced triangulation. The strategy for constructing $\tilde{f}$ is to construct (continuous) maps $\tilde{f}^{(i)}$ from the i-skeleta $\widetilde{M}^{(i)}$ of $\widetilde{M}$ to T for $i = 0, 1, 2, 3$; each $\tilde{f}^{(i)}$ will be $\pi_1(M)$-equivariant, and $\tilde{f}^{(i)}$ will extend $\tilde{f}_{i-1}$ for $i = 1, 2, 3$. Of course we'll define $\tilde{f}$ to be $\tilde{f}_3$.

To construct $\tilde{f}^{(0)}$ we first pick a *complete system of orbit representatives* for the action of $\pi_1(M)$ on $M^{(0)}$, i.e., a set $S^{(0)} \subset M^{(0)}$ such every orbit for the action of $\pi_1(M)$ on $M^{(0)}$ meets $S^{(0)}$ in exactly one point. Now, using the fact that $\pi_1(M)$ acts freely on M – and hence on $M^{(0)}$ – we can see that if $h^{(0)}$ is any map whatever from $S^{(0)}$ to the vertex set $T^{(0)}$ of T then $h^{(0)}$ has one and only one extension $\tilde{f}^{(0)}$: $\widetilde{M}^{(0)}$ which is $\pi_1(M)$-equivariant. Uniqueness is clear, since any such map must in particular satisfy

$$\tilde{f}^{(0)}(\gamma \cdot s) = \gamma \cdot h^{(0)}(s) \quad \text{for all } s \in S^{(0)} \text{ and } \gamma \in \pi_1(M). \tag{2.1.1}$$

To get existence we notice that (2.1.1) makes sense as a definition of $\tilde{f}^{(0)}$ because every vertex in $M^{(0)}$ can be expressed *in exactly one way* in the form $\gamma \cdot s$ with $\gamma \in \pi_1(M)$. The point here is that if $\gamma \cdot s = \gamma' \cdot s$ then s is a fixed point of $\gamma^{-1}\gamma'$, and since the action is free we must have $\gamma = \gamma'$. Now that we know Definition 2.1.1 makes sense, there is no problem checking that the map $\tilde{f}^{(0)}$ is equivariant.

It's significant that we could have started with *any* map $h : S^{(0)} \to T^{(0)}$ for this step. (Actually at this stage we don't even need h to map $S^{(0)}$ into $T^{(0)}$, but that will be nice to know later.) This illustrates how far our construction is from being canonical. The flexibility in the definition of $\tilde{f}$ turns out to be very useful; I'll return to this issue later in this section.

Now suppose that $\tilde{f}^{(i)}$ has been constructed for a given i with $0 \leqslant i \leqslant 2$. Let us pick a complete system of orbit representatives $S^{(i+1)}$ for the action of $\pi_1(M)$ on the set of all $(i+1)$-simplices of $\widetilde{M}$. For any simplex $\sigma \in S^{(i+1)}$ we have a map $\tilde{f}^{(i)}|\partial\sigma : \partial\sigma \to T$. Since T is contractible, this map can be extended to a map $\tilde{f}_\sigma : \sigma \to T$. Now there is a unique continuous, $\pi_1(M)$-equivariant map $\tilde{f}^{(i+1)} : M^{(i+1)} \to T$ which restricts to h_σ on each $\sigma \in S^{(i+1)}$. Indeed, such a map must obviously be given by

$$\tilde{f}^{(i+1)}(\gamma \cdot x) = \gamma \cdot h_\tau(x) \quad \text{for all } \tau \in S^{(0)},\ x \in \tau \text{ and } \gamma \in \pi_1(M).$$

We can show, almost exactly as in the construction of $\tilde{f}^{(0)}$, that the map given by this formula is well-defined and equivariant. Continuity is then easy.

It's easy to adapt this construction so as to guarantee that the map $\tilde{f}$ is simplicial with respect to some $\pi_1(M)$-invariant triangulation of $\widetilde{M}$ (and a given triangulation of T). All that we have to notice is that the in the induction step, the map $\tilde{f}^{(i)}|\partial\sigma$ can be extended to a *simplicial* map $\tilde{f}_\sigma : \sigma \to T$; this follows from the simplicial approximation theorem for pairs (see [62, Chapter 3, Section 4, Theorem 8 and Lemma 1]).

2.2. *Constructing a surface*

Now, fixing a $\pi_1(M)$-invariant triangulation of $\widetilde{M}$ in which $\tilde{f}$ is simplicial, let's consider a point x of T which is not a vertex. Consider the subset $P = \tilde{f}^{-1}(x)$ of $\widetilde{M}$, and for any i-simplex σ of $\widetilde{M}$ consider the subset $P \cap \sigma$ of σ. If $\tilde{f}$ does not map σ onto the edge e of T containing x then $P \cap \sigma = \emptyset$. (This is always the case if $i = 0$.) If $\tilde{f}$ does map σ onto e, then since no vertex of σ is mapped to x, the set $P \cap \sigma$ is an $(i-1)$-cell properly embedded in σ: if we think of the simplex σ as being embedded in an affine space, $P \cap \sigma$ is the intersection of σ with a hyperplane missing the vertices of σ. Now since P meets every simplex of $\widetilde{M}$ either in the empty set or in a properly embedded codimension-1 cell, it's easy to see that P is a 2-manifold, properly embedded in $\widetilde{M}$.

(The most interesting point is checking that P is locally Euclidean at any point z where it meets a 1-simplex τ of $\widetilde{M}$. The 2-simplices incident to τ look like the pages of a cyclic book with τ as binding. The set P meets each page in a 1-cell which has an endpoint at z and is otherwise disjoint from τ. Since P meets each 3-simplex in a 2-cell, we can recover its whole intersection with the open star of τ by connecting every two successive 1-cells with a 2-cell, giving an open 2-disk.)

The basic idea for associating surfaces in M with actions of $\pi_1(M)$ on trees is now easy to explain. Let's denote by E the set of all midpoints of edges of the tree T. Since, by the discussion above, $\tilde{f}^{-1}(x)$ is a properly embedded 2-manifold in $\widetilde{M}$ for each $x \in E$, the set $\widetilde{F} = \tilde{f}^{-1}(E)$ is a properly embedded 2-manifold in $\widetilde{M}$. On the other hand, since E is clearly $\pi_1(M)$-invariant and $\tilde{f}$ is $\pi_1(M)$-equivariant, $\widetilde{F}$ is invariant under the action of $\pi_1(M)$ on $\widetilde{M}$. So $\widetilde{F}$ is the inverse image, under the covering projection, of some properly embedded 2-manifold $F \subset M$.

In Section 2.4 I will show how to modify the map $\tilde{f}$ so that the surface $F \subset M$ that it defines is essential. Before doing this I need to deal with a small technical point, and make some definitions and remarks.

The technical point involves a slightly stronger version of the condition that $\widetilde{F}$ be a 2-manifold, which makes life much simpler when we are worrying about such things as making F essential. Let $E \subset T$ be a discrete set containing no vertices of T. A continuous map $\tilde{f} : \widetilde{M} \to T$ is said to be *transverse to* X if each point $z \in \tilde{f}^{-1}(X)$ has a neighborhood U with a homeomorphism $h : U \to V \times (0, 1)$, for some open subset V of $\mathbf{R}^2$, such that $\tilde{f}|U = j \circ q \circ h$, where $q : (0, 1) \times (0, 1) \to (0, 1)$ is the projection to the second factor, and j is some homeomorphism of $(0, 1)$ onto an open interval in some edge of T. It's immediate that if $\tilde{f}$ is transverse to X then $\tilde{f}^{-1}(X)$ is a bicollared 2-manifold in $\widetilde{M}$. On the other hand, if you examine the proof that I gave above that, for the equivariant map $\tilde{f}$

that I constructed, $\tilde{f}^{-1}(E)$ is a properly embedded 2-manifold, you will have no trouble obtaining the stronger conclusion that this map $\tilde{f}$ is transverse to E.

It follows from the discussion I gave above that if $\tilde{f}: \widetilde{M} \to T$ is a $\pi_1(M)$-equivariant map transverse to E, then $\tilde{f}^{-1}(E)$ is the inverse image under the covering projection of a properly embedded surface F in M. This surface is well-defined once we have fixed the map $\tilde{f}$. I'll say that a bicollared surface $F \subset M$ is *dual to* the given action of $\pi_1(M)$ on T if it arises via this construction from some $\pi_1(M)$-equivariant map transverse to E.

2.3. *Remarks on dual surfaces*

When the fundamental group of a compact, orientable, irreducible 3-manifold M acts simplicially on a tree T, properties of the action are reflected in the behavior of surfaces dual to the action. The following statement covers a lot of applications:

2.3.1. *If F is a dual surface to an action of $\pi_1(M)$ on a tree T, then*:
- (i) *for each component C_i of $M - F$, the subgroup* $\mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ *of $\pi_1(M)$ is contained in the stabilizer of some vertex of T; and*
- (ii) *for each component F_j of F, the subgroup* $\mathrm{im}(\pi_1(F_j) \to \pi_1(M))$ *of $\pi_1(M)$ is contained in the stabilizer of some edge of T.*

(Of course, as I mentioned in Section 1.2, the subgroup $\mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ is defined only up to conjugacy in $\pi_1(M)$, but statement (i) makes sense because the conjugate of a vertex stabilizer is still the stabilizer of a (probably different) vertex. Likewise for (ii).)

Properties 2.3.1(i) and (ii) are just about immediate from the definition of a dual surface. If F is defined by an equivariant map $\tilde{f}: \widetilde{M} \to T$, transverse to the set E of midpoints of edges of T, then for any component C_i of $M - F$, the subgroup $\Gamma_i = \mathrm{im}(\pi_1(C_i) \to \pi_1(M))$ is the stabilizer of a component $\widetilde{C}$ of $M - \widetilde{F}$, where $F = f^{-1}(E)$. (Varying Γ_i within its conjugacy class just replaces $\widetilde{C}$ by another component of $M - \widetilde{F}$.) Now $\widetilde{F}$ maps $\widetilde{C}$ into a component S of $T - T^0$, where T^0 denotes the set of vertices of T. The equivariance of $\widetilde{F}$ implies that Γ_0 stabilizes S. But S is just the open star, relative to the first barycentric subdivision of T, of a vertex s of T; and since the action of $\pi_1(M)$ is simplicial, the stabilizer of S is the same as the stabilizer of s. This proves (i), and (ii) is even easier. □

By the way, if you feel that the inclusions of subgroups given by 2.3.1 really ought to be equalities, you should have a look at the end of this section.

If we are in the really stupid case where the dual surface F is *empty*, then the only component of $M - F$ is the whole manifold M, and it follows from 2.3.1(ii) that $\pi_1(M)$ fixes a vertex of T. Remember that we express this by saying that the action of $\pi_1(M)$ on T is trivial. So:

2.3.2. *If $\pi_1(M)$ acts nontrivially on T then any surface dual to the action is nonempty.*

Sometimes it's useful to think of a dual surface $F \subset M$ in terms of a map of M into a graph, i.e., a 1-dimensional CW-complex. In fact, since Γ is assumed to act on T without inversions, the orbit space $\mathcal{G} = T/\Gamma$ has the structure of a graph in a natural way, the open 1-cells being the homeomorphic images of edges of T. (The presence of inversions would make certain edges get folded in two.) Now if $\tilde{f} : \widetilde{M} \to T$ is a $\pi_1(M)$-equivariant map transverse to E, there is a unique map $f : M \to \mathcal{G}$ such that the diagram

$$\begin{array}{ccc} \widetilde{M} & \xrightarrow{\tilde{f}} & T \\ \downarrow & & \downarrow \\ M & \xrightarrow{f} & \mathcal{G} \end{array}$$

commutes. We clearly have $F = f^{-1}(\overline{E})$, where $\overline{E}$ is the image of E in $\mathcal{G}$; you may think of E as the set of midpoints of edges of $\mathcal{G}$. If there were a simple characterization of maps $f : M \to \mathcal{G}$ that are induced by equivariant maps from $\widetilde{M}$ to T, this would give a simplified definition of dual surfaces, but the fact is that characterizing such maps is pretty messy. Still, this way of looking at a dual surface is sometimes useful. For one thing, since f is itself transverse to $\overline{E}$ (in essentially the same sense as before), and since each point of $\overline{E}$ is two-sided in $\mathcal{G}$, it's an easy exercise to conclude that each component of F is two-sided in M. So:

2.3.3. *If $\pi_1(M)$ acts on T without inversions then the components of any dual surface are two-sided.*

(Since we've assumed M to be orientable, 2.3.3 is the same as saying that F is orientable, but 2.3.3 would be true even without this assumption.)

By and large, the dual surfaces that I will be working with in this chapter will be piecewise linear with respect to some triangulation of the ambient manifold. As I mentioned at the end of Section 1.4, a two-sided, properly embedded PL hypersurface in a PL manifold is always bicollared. However, one can show without ever mentioning a PL structure on M, or putting any additional restrictions on the surface, that:

2.3.4. *Any surface that's dual to an action of $\pi_1(M)$ on a tree is bicollared.*

The point is that the condition that $f : M \to \mathcal{G}$ is transverse to $\overline{E}$ immediately implies that the surface $F = f^{-1}(\overline{E}) = \tilde{f}^{-1}(E)$ is *locally flat:* this simply means that every point $x \in \widetilde{F}$ has a neighborhood V in $\widetilde{M}$ which can be mapped homeomorphically onto $\mathbf{R}^3$ in such a way that $\widetilde{F} \cap V$ is mapped onto $\mathbf{R}^2$. And it is a theorem of Morton Brown's [12] that a locally flat, two-sided, properly embedded hypersurface in an n-manifold is always bicollared.

Now let's turn to the question of how the notion of dual surface is related to the construction described in Section 1.4 that associates a tree with a surface. If F is a bicollared surface in M, and if T is the tree, on which $\pi_1(M)$ acts without inversions, which is given by the construction of 1.4, then since T is by definition a quotient of the universal covering $\widetilde{M}$, we have a natural map $\widetilde{M} \to T$. It's straightforward to check that this map is equivariant and transverse to the midpoints of edges of T, and that F is the dual surface it defines.

The situation is more subtle if we begin with an action of $\pi_1(M)$ without inversions on a tree T, choose a dual surface F to the action, and compare the given action of $\pi_1(M)$ on T with its action on the tree T' which is associated with F by the construction of Section 1.4. This construction does make sense here, since we just saw that F is bicollared. It's natural to wonder whether T' and T are equivariantly isomorphic. This would mean that the inclusions given by 2.3.1 were equalities.

But this is false. In fact, there are perfectly reasonable examples in which the fundamental group of a compact, irreducible, orientable 3-manifold M acts on a tree T in such a way that the stabilizers of all the edges and vertices of T in $\pi_1(M)$ are infinitely generated. This means that we can never have equality in 2.3.1(i) or (ii), regardless of which dual surface we choose.

We can see this phenomenon by using an especially simple example of a tree: the real line $\mathbf{R}$, triangulated with a vertex at each integer point. The action of the group $\mathbf{Z}$ on $\mathbf{R}$ by translations is a simplicial action. Now if Γ is any group and $\phi : \Gamma \to \mathbf{Z}$ is a homomorphism, pulling back the standard action of $\mathbf{Z}$ on $\mathbf{R}$ via ϕ gives an action of Γ on $\mathbf{R}$, which is nontrivial if and only if the homomorphism ϕ is nontrivial. The stabilizer of any edge or vertex is simply the kernel of ϕ. Now Stallings's fibration theorem [34,63] asserts that if M is a compact, irreducible, orientable 3-manifold, and $\phi : \pi_1(M) \to \mathbf{Z}$ is a homomorphism, then $K = \ker\phi$ can be finitely generated only if M is a locally trivial fiber bundle over S^1 and ϕ is the homomorphism induced by the bundle map. In this case K is just the image of the fundamental group of a fiber F, and it's easy to see that F is just a dual surface; this gives a class of examples where we do have $T' = T$.

In general, however, there is no reason why a homomorphism $\pi_1(M) \to \mathbf{Z}$ should be realized by a bundle map to S^1. Among 3-manifolds which have positive first Betti number, and whose fundamental groups therefore admit homomorphisms onto $\mathbf{Z}$, there are certainly plenty of examples – my unjustified intuition says they are a majority – which cannot be fibered over S^1 at all. You can get a sense of the issues involved by looking at [66].

Still, it is not hard to understand the relationship between the trees T and T' in general. By fiddling a little you should be able to show that with a little care in choosing the bicollaring of the dual surface F which is used in constructing the tree T', we can guarantee that the equivariant map $\tilde{f} : \widetilde{M} \to T$, which defined F, factors through a $\pi_1(M)$-equivariant map $T' \to T$. Although I won't be using this equivariant map in this chapter, it certainly gives a nice picture. For example, you can read off the inclusions 2.3.1(i) and (ii) from the existence of this map, since equivariance implies that every edge or vertex stabilizer of T' is contained in an edge or vertex stabilizer of T. This is presumably the way that mathematical aristocrats – to borrow a phrase of Raoul Bott's – think about 2.3.1. Personally I have the soul of a petit bourgeois. Otherwise I would never have finished this chapter by the deadline.

2.4. *Making a dual surface essential*

Now I'll indicate how, given a compact, irreducible 3-manifold M and an action of $\pi_1(M)$ on a tree T, you can construct an essential surface in M that's dual to the given action. We saw in the last section that there is some bicollared surface dual to the action. What I'll give here is a construction that can be carried out whenever a given surface F, dual to the action, is not essential. This construction replaces F by a new surface F', also dual to the action. I will then point out that F' is always "simpler" than F in a suitable sense that I'll make precise; this will imply that by repeating the construction a finite number of times we always arrive at an essential surface.

The most important part of the definition of an essential surface is the π_1-injectivity condition 1.5.1(i). The heart of the matter is therefore to give a construction for simplifying F in the case where 1.5.1(i) fails to hold; by the discussion following Definition 1.5.1, this is the case in which there is a compressing disk D for F. In this situation there is a natural operation, called a *compression*, by which you can replace F by a new surface F': you remove from F an annular neighborhood A of the simple closed curve ∂D, and to the resulting surface you attach two parallel copies of D, say D_1 and D_2, whose boundaries are the two components of ∂A. (Defining "parallel" here involves part (c) of the definition of a compressing disk given in Section 1.5, and it's easy to show that F' is bicollared.) A little later I will point out a precise and useful sense in which F' is "simpler" than F. Right now let me point out why the surface F' is dual to the action of $\pi_1(M)$ on T.

Since F is dual to the action, there is a $\pi_1(M)$-equivariant map $\tilde{f}:\tilde{M} \to T$, transverse to E, such that $f^{-1}(E) = p^{-1}(F)$, where $p:\tilde{M} \to M$ is the covering projection. We are required to find another $\pi_1(M)$-equivariant map $\tilde{f}':\tilde{M} \to T$, transverse to E, such that $(f')^{-1}(E) = p^{-1}(F)$. Let's choose a nice neighborhood B of D in M, so that B is homeomorphic to a ball and meets F in the annulus A, which is properly embedded in B. Then B is the union of a ball X^+ and a solid torus X^-, where $X^+ \cap X^- = A$; we may take the disks D_1 and D_2 to be contained in X^+, and properly embedded in X^-. (Formal proofs of things like this can be given by using regular neighborhood theory. See [35].)

Now let $\tilde{B}$ be a component of $p^{-1}(B)$, so that p maps $\tilde{B}$ homeomorphically onto B, and let $\tilde{A}$, $\tilde{D}_1$, $\tilde{D}_2$, $\tilde{X}^+$ and $\tilde{X}^-$ denote the inverse images in $\tilde{B}$ of A, D_1, D_2, X^+ and X^-. Since $\tilde{f}$ is transverse to E and $\tilde{A} = \tilde{B} \cap f^{-1}(E)$, and since T is a tree, $\tilde{f}$ must map X^+ and X^- to the closures of different components of $T - E$, say Y^+ and Y^-. Let h denote the map from $\partial\tilde{B} \cup D_1 \cup D_2$ to T which agrees with $\tilde{f}$ on $\partial\tilde{B}$ and maps D_1 and D_2 to the point $p(A)$ of E. Then h can be extended to a map $g:\tilde{B} \to T$ such that $g^{-1}(E) = D_1 \cup D_2$. To see this, note that D_1 and D_2 divide B into three balls. If C is any of these balls, h maps ∂C to the closure of either Y^+ or Y^-; this makes it possible to extend $h|\partial C$ to a map from C to the closure or Y^+ or Y^- in such a way that the interior of C is mapped to either Y^+ or Y^-.

Now g admits a unique extension to a $\pi_1(M)$-equivariant map $\tilde{f}'_B : p^{-1}(B) \to T$, since each component of $p^{-1}(B)$ is the image of $\tilde{B}$ under a unique element of $\pi_1(M)$. If we define $\tilde{f}'$ to agree with f'_B on its domain and with $\tilde{f}$ on the rest of $\tilde{M}$, it is clear that $\tilde{f}'$ is continuous and $\pi_1(M)$-equivariant, and that $(\tilde{f}')^{-1}(E) = p^{-1}(F')$. You should convince yourself that with just a little more care in the construction of the extension g of h, we can guarantee that $\tilde{f}'$ is actually transverse to E. This shows that the surface F' is dual to the action.

(One comment that's worth making about the map $\tilde{f}'$ that's been constructed here is that even if $\tilde{f}$ were simplicial with respect to some subdivision of $\tilde{M}$ and the given triangulation of T – as was the case with the map $\tilde{f}$ that I constructed in Section 2.1 – it's almost certainly necessary to subdivide T before $\tilde{f}'$ can be made simplicial.)

In the other cases where F fails to be essential, the construction of F' is significantly easier. We saw in Section 2.3 that F can't fail to satisfy condition 1.5.1(iv). If F satisfies condition (i) of 1.5.1 but does not satisfy both conditions (ii) and (iii), it has a component F_0 which is either a 2-sphere or a boundary-parallel surface. If F_0 is a 2-sphere, then since M is irreducible, F_0 bounds a ball B_0; we may assume F_0 to be chosen so that B_0 doesn't contain any other 2-sphere component of F. Since F already satisfies 1.5.1(i), this implies that B_0 contains no component of F whatever.

It's not hard to see that $F' = F - F_0$ is again a dual surface. For example, in the case where F_0 is boundary-parallel, there's a submanifold M' which is a deformation-retract of M, such that $M' \cap F = F'$. If we think of a deformation-retraction from M to M' as a map $\rho : M \to M$, then ρ is covered by a $\pi_1(M)$-equivariant map $\tilde{\rho} : \widetilde{M} \to \widetilde{M}$. If $\tilde{f} : \widetilde{M} \to T$ is a $\pi_1(M)$-equivariant map, transverse to E, with $\tilde{f}^{-1}(E) = p^{-1}(F)$, then $\tilde{f}' = \tilde{f} \circ \tilde{\rho}$ is also $\pi_1(M)$-equivariant map and transverse to E, and we have $\tilde{f}^{-1}(E) = p^{-1}(F')$. I'll let you work out the case where F_0 is a 2-sphere.

Now we need to address the sense in which F' is "simpler" than F. For any compact 2-manifold F, let me define the *complexity* of F to be the nonnegative integer

$$c(F) = \sum \bigl(2 - \chi(F_i)\bigr)^2, \tag{2.4.1}$$

where F_i ranges over the components of F and χ denotes the Euler characteristic. Since a compact, connected 2-manifold has Euler characteristic at most 2, the expressions whose squares appear in the sum (2.4.1) are always nonnegative.

The reason this complexity is useful is that the operations I've described above, which replace a nonessential dual surface by a new dual surface, usually decrease the complexity. Specifically, this is always true of the first operation I described, the compression, which can be carried out when the given dual surface F fails to satisfy condition (i) of Definition 1.5.1. To see this, notice that if F' is obtained by a compression from F, we have $F' = (F - \operatorname{int} A) \cup D_1 \cup D_2$, where A is an annulus in F whose core curve doesn't bound a disk in F, and D_1 and D_2 are disks with $(D_1 \cup D_2) \cap F = \partial D_1 \cup \partial D_2 = \partial A$. If F_0 denotes the component of F containing A, then F' is obtained from F by replacing F_0 by the 2-manifold $F_0' = (F_0 - \operatorname{int} A) \cup D_1 \cup D_2$, which has either one or two components. Now when we form the union of two surfaces that meet along a collection of common boundary curves, the Euler characteristic of the union is the sum of the Euler characteristics of the two pieces; hence

$$\chi(F_0') = \bigl(\chi(F_0) - \chi(A)\bigr) + \bigl(\chi(D_1) + \chi(D_2)\bigr) = \chi(F_0) + 2.$$

Now there are two cases. If F_0' is connected, the effect of our operation on the sum (2.4.1) is to replace the term $(2 - \chi(F_0))^2$ by the term the term $(2 - \chi(F_0'))^2$, where $\chi(F_0') = \chi(F_0) + 2 \leqslant 2$, so it's clear that $c(F') < c(F)$.

If F_0' has two components, say F_α' and F_β', then neither F_α' nor F_β' is a sphere; this is because the core curve of A did not bound a disk in F. Hence if we set $a = 2 - \chi(F_\alpha')$ and $b = 2 - \chi(F_\beta')$, then a and b are both strictly positive. Now we have

$$2 - \chi(F_0) = 4 - \chi(F_0') = 4 - \bigl(\chi(F_\alpha') + \chi(F_\beta') - 2\bigr) = a + b.$$

So the effect of our operation on the sum (2.4.1) is to replace the term $(a+b)^2$ by the term a^2+b^2. Since a and b are strictly positive, we again have $c(F') < c(F)$.

The other operations on F that I described, for the cases where condition (ii) or (iii) of Definition 1.5.1 fails, amount to discarding a component. This cannot increase the complexity of F, although it may keep it the same if the component in question is a sphere. It's now pretty clear how to get an essential surface which is dual to the given action. We choose a surface F whose complexity is minimal among all dual surfaces; and among those dual surfaces having minimal complexity, we choose F so as to have the smallest possible number of components. If F were not essential, one of our operations would produce either a dual surface of strictly smaller complexity, or one having the same complexity but fewer components, and in either case we'd have a contradiction.

2.5. *Applications, I: Nonseparating surfaces*

The simplest consequence of the constructions described in this section is that if M is a compact, orientable, irreducible 3-manifold such that $\pi_1(M)$ admits a nontrivial action on a tree, then M contains an essential surface. Conversely according to Proposition 1.5.2, if M contains an essential surface then $\pi_1(M)$ admits a nontrivial action on a tree.

In Section 2.3 I already mentioned an especially simple way of constructing an action of $\pi_1(M)$ on a tree: pull back the action of $\mathbf{Z}$ on $\mathbf{R}$ by translations, via some homomorphism from $\pi_1(M)$ to $\mathbf{Z}$. Such homomorphisms are in canonical bijective correspondence with elements of $H^1(M;\mathbf{Z})$. So the general construction described in this section allows one to associate (noncanonically) an essential surface F with any nontrivial element $c \in H^1(M;\mathbf{Z})$. It's a good exercise to show that the image in $H_1(M, \partial M;\mathbf{Z})$ of the fundamental class $[F]$ of F is the Poincaré dual of c. (To make this precise, we need to orient the components of F so that $[F]$ will be well defined; choosing the right orientations is part of the exercise. Here's a hint: the equivariant map $\tilde{f}:\widetilde{M} \to \mathbf{R}$ from which F is constructed induces a map from M to S^1.)

Since $[F] \neq 0$, there is at least one component F_0 of F such that $[F_0] \neq 0$, which is tantamount to saying that F does not separate M. In particular:

PROPOSITION 2.5.1. *If M is a compact, orientable, irreducible 3-manifold with positive first Betti number, then M contains a nonseparating essential surface.*

This particular result is fundamental in 3-manifold theory. A *Haken manifold* is defined to be a compact, orientable, irreducible 3-manifold which is either homeomorphic to a ball or contains an essential surface. If M is a compact, orientable, irreducible 3-manifold with nonempty boundary, then either ∂M has a 2-sphere component, in which the irreducibility of M implies that M is a ball, or some component of ∂M has genus > 0, in which case an elementary application of Poincaré duality (see, for example, the proof of Lemma 4.9 in [34]) shows that M has positive first Betti number, so that M contains an incompressible surface according to Proposition 2.5.1. So we may state the

COROLLARY 2.5.2. *Every compact, orientable, irreducible 3-manifold with nonempty boundary is a Haken manifold.*

Now suppose that M is a Haken manifold not homeomorphic to a ball. If we choose an essential surface $F \subset M$, then splitting M along F gives a new – possibly disconnected – manifold M', each component of which has nonempty boundary, and it's not hard to conclude from the irreducibility of M and the essentiality of F that each component of M' is irreducible. So each component of M' is a Haken manifold. If some component of M' is not a ball, it follows that there is an essential surface F' in some component of M', and we can split M' along F' to get a manifold M'', each component of which is a Haken manifold. It was first shown by Haken – see [34, pp. 140–142] for a simplified proof – that if the surfaces F' are chosen with a little care then this process must terminate; that is, there is some $n \geqslant 0$ such that every component of $M^{(n)}$ is a ball. The finite sequence $M, M', \ldots, M^{(n)}$ is called a *hierarchy* for the manifold M.

Some of the deepest results in 3-manifold theory are theorems about Haken manifolds that are proved by induction on the length of a hierarchy. To prove such a theorem we need only prove that it is true for a given manifold M whenever it is true for each component of the manifold M' obtained by splitting M along an essential surface. The first results of this type were obtained by Waldhausen; for the special case of a closed Haken manifold M his results imply that every irreducible 3-manifold M_1 with $\pi_1(M_1) \cong \pi_1(M)$ is homeomorphic to M, that every outer automorphism of $\pi_1(M)$ is induced by a self-homeomorphism of M, and that self-homeorphisms of M which induce the same outer automorphism are isotopic. Another particularly famous example is Thurston's theorem characterizing those Haken manifolds which admit hyperbolic metrics of finite volume.

Corollary 2.5.2 applies in particular when M is the exterior of a tame knot in a closed, orientable 3-manifold Σ. By a *knot* in M I mean a subset K homeomorphic to S^1; to say that M is tame means that it has a tubular neighborhood – that is, a neighborhood V which can be mapped homeomorphically to $S^1 \times D^2$ in such a way that K is mapped onto $S^1 \times \{0\}$. (Every smooth or piecewise linear knot is automatically tame.) The *exterior* of K is the closure of $\Sigma - V$. If Σ is a homology 3-sphere (for example S^3), there is a stronger version of Corollary 2.5.2. The proof of the corollary gave a nonseparating essential surface F in M, but I claim we can take F to have a connected boundary. This is an easy consequence of a relative version of the main construction of this section; since I'll need the relative version elsewhere in the chapter, I will spell out the conclusion:

PROPOSITION 2.5.3. *Let M be a compact, orientable, irreducible 3-manifold, and suppose that we are given a nontrivial action of $\pi_1(M)$ on a tree T. Let $B_1, \ldots, B_k \subset \partial M$ be disjoint (compact) subpolyhedra, and consider the action of each $\pi_1(B_i)$ on T obtained by pulling back the action of $\pi_1(M)$ via the inclusion homomorphism $\pi_1(B_i) \to \pi_1(M)$. (These actions are well-defined up to equivalence; see Section 1.2.) Suppose that for $i = 1, \ldots, k$ we are given a $\pi_1(B_i)$-equivariant map $\tilde{g}_i : \widetilde{B}_i \to T$, where $\widetilde{B}_i$ denotes the universal covering space of B, and suppose that each $\tilde{g}_i$ is transverse to the set E of all midpoints of edges of T, so that $\tilde{g}_i^{-1}(E)$ is the pre-image of a unique closed 1-manifold $C_i \subset B$. Then there is a dual surface F to the action of $\pi_1(M)$ on T such that $B_i \cap \partial F \subset C_i$ for $i = 1, \ldots, k$.*

After a few preliminaries, proving this is just a matter of carrying out the constructions described in Sections 2.1, 2.2 and 2.4 with a little care. Working out the details is a perfect exercise in understanding these constructions. One uses the equivariance property of $\tilde{g}_i$ to show that $\tilde{g}_i$ factors through a map $\tilde{g}'_i : \widetilde{B}_i/N_i$ to T, where N_i is the kernel of the inclusion homomorphism $\pi_1(B_i) \to \pi_1(M)$. The quotient surface $\widetilde{B}_i/N_i$ can be identified with a boundary component of the universal cover $\widetilde{M}$ of M. Let's define $\tilde{g}' : (\widetilde{B}_1/N_1) \cup \cdots \cup (\widetilde{B}_k/N_k) \to T$ to be the map which restricts to $\tilde{g}'_i$ on each $\widetilde{B}_i/N_i$. There is a unique extension of $\tilde{g}$ to a $\pi_1(M)$-equivariant map $\tilde{f}_\partial : p^{-1}(B) \to T$, where $p : \widetilde{M} \to M$ is the covering projection and $B = B_1 \cup \cdots \cup B_k$.

If one begins with a triangulation of M in which B is a subcomplex, then in carrying out the inductive construction given in Section 2.1 for the $\pi_1(M)$-equivariant map $\tilde{f}$, simplicial with respect to some $\pi_1(M)$-equivariant subdivision of $\widetilde{M}$ and the given triangulation of T, one can make the choices of extensions $\tilde{f}_\sigma$ in such a way that for each simplex $\sigma \subset p^{-1}(B)$ we have $\tilde{f}_\sigma = \tilde{f}_\partial|\sigma$. According to the discussion in Sections 2.2 and 2.3, $\tilde{f}$ is transverse to $E \subset T$. The dual surface defined by $\tilde{f}$ – let's call it F_0 – has the property that $B_i \cap F_0 = C_i$ for $i = 1, \ldots, k$. The arguments of Section 2.4 show that by a finite sequence of modifications we can replace F_0 by an essential surface dual to the action. If you examine the effect of these operations on the boundary, you will find that $\partial F \subset \partial F_0$, so that $B_i \cap \partial F \subset C_i$ for $i = 1, \ldots, k$, as asserted in Proposition 2.5.3.

Now, as I was saying, Proposition 2.5.3 can be used to show that if M is the exterior of a knot K in a homology 3-sphere Σ, then there is an essential nonseparating surface $F \subset M$ with connected boundary. To see this, first notice that $H_1(M; \mathbf{Z})$, which is the abelianization of $\pi_1(M)$, is infinite cyclic and generated by the meridian class, as is easily deduced from the Mayer–Vietoris theorem. Hence, up to sign, there's a unique homomorphism $\phi : \pi_1(M) \to \mathbf{Z}$, and ϕ maps the conjugacy class λ defined by the longitude of K to 0, and maps the conjugacy class μ defined by the meridian of K to a generator of $\mathbf{Z}$. (See Boyer's chapter for terminology concerning meridians and longitudes.) If we identify ∂M with $S^1 \times S^1$ so that $S^1 \times \{\text{point}\}$ is a meridian and $\{\text{point}\} \times S^1$ is a longitude, then the universal cover of ∂M becomes identified with $\mathbf{R} \times \mathbf{R}$. The hypotheses of Proposition 2.5.3 now hold if we set $T = \mathbf{R}$ and $B = \partial M$, let $\pi_1(M)$ act on $\mathbf{R}$ by the pullback via ϕ of the action of $\mathbf{Z}$ on $\mathbf{R}$ by translations, and define $\tilde{g} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ to be the projection to the first factor. The 1-manifold C described in the statement of Proposition 2.5.3 has the form $\{\text{point}\} \times S^1$ and is therefore a longitude. Hence proposition gives a dual surface F whose boundary is either a longitude or the empty set. But if $\partial F = \emptyset$ then the equivariant map defining F maps the boundary of $\widetilde{M}$ into the complement of the set E of midpoints of edges of $\mathbf{R}$, hence into an interval $[n - \frac{1}{2}, n + \frac{1}{2}]$; by equivariance, the vertex n of $\mathbf{R}$ is fixed by $\pi_1(\partial M)$. This is absoid, since μ acts on $\mathbf{R}$ by a unit translation.

So we do indeed have an essential surface $F \subset M$ whose boundary is connected, and is in fact a longitude. It's easy to see that in the tubular neighborhood V of K whose interior was removed from Σ to get M, there is an annulus A whose boundary curves are the knot K and a longitude, which we can take to be ∂F. So $F^+ = F \cup A$ is an embedded (but not properly embedded) compact, orientable surface in Σ with boundary F. Such a surface is called a Seifert surface. While Seifert surfaces are great for making plastic models, properly embedded surfaces in knot exteriors are generally more useful for theoretical work, as you will quickly discover if you try to write down the properties of F^+ that translate the condition that F is essential. Of course, F is so closely related to F^+ that one quickly slips into the habit of calling F a Seifert surface.

One immediate consequence of the existence of an essential Seifert surface in a knot K is that if K is nontrivial then the *group* of K, defined to be $\pi_1(\Sigma - K) \cong \pi_1(M)$, has a nonabelian free subgroup. (To say that K is nontrivial means that it doesn't bound a disk in Σ. So a Seifert surface, which by definition is orientable and has a connected boundary, must have genus > 0, so that $\pi_1(F)$, which injects into $\pi_1(M)$ by condition (i) of Definition 1.5.1, is a nonabelian free group.) This is a simple illustration of how the study of essential surfaces gives information about the structure of a knot group. I'll give a fancier illustration of this in Sections 5.6 and 6.

2.6. *Applications, II: Free products and stuff*

It's routine to show that if an irreducible 3-manifold M contains an essential (bicollared) disk D then $\pi_1(M)$ is either a nontrivial free product or an infinite cyclic group. If D separates M then the irreducibility of M can be used to show that neither component of $M - D$ is simply connected, so that M is a nontrivial free product by van Kampen's theorem. If D doesn't separate M, van Kampen's theorem exhibits $\pi_1(M)$ as a free product of $\mathbf{Z}$ with $\pi_1(M - D)$, which may or may not be trivial; in either case we get the desired conclusion.

Conversely, if $\pi_1(M)$ is either a nontrivial free product or an infinite cyclic group, then M contains an essential disk. To prove this from the point of view of this section, the main thing you have to notice is that a group Γ which is either infinite cyclic or a free product admits a nontrivial action on a tree T in which the stabilizer of each edge is trivial. If Γ is infinite cyclic we can take $T = \mathbf{R}$. If Γ is a nontrivial free product, one can either construct the tree from the algebra of a free product – a good exercise – or describe it topologically as follows. Think of Γ as the fundamental group of a space X which is the union of two disjoint nonsimply connected spaces A and B and a topological arc which meets each of A and B in a single point. If $\widetilde{X}$ is the universal cover of X and $p : \widetilde{X} \to X$ is the covering projection, we can obtain T as a quotient of $\widetilde{X}$ by identifying each component of $p^{-1}(A \cup B)$ to a point. The usual action of $\pi_1(X)$ on $\widetilde{X}$ induces an action on T with the required property.

Now if $\pi_1(M)$ acts on a tree with trivial edge stabilizers, then according to the results of Sections 2.2 and 2.4, there is an essential dual surface F to the action; and according to 2.3.1(ii), the fundamental group of each component of F is contained in the stabilizer of an edge of T and is therefore trivial. As my definition of essential surface rules out 2-sphere components, the components of F must be essential disks.

There is a useful generalization of the notion of a free product. Suppose we are given groups A, B and C, and injective homomorphisms $i : C \to A$ and $j : C \to B$. The *free product of A and B amalgamated over C*, denoted rather vaguely by $A \star_C B$, is defined to be quotient of the free product $A \star B$ obtained by adding the relations $i(c) = j(c)$ for all $c \in C$. It is a theorem, of which you will find an elegant account in [40], that the natural homomorphisms from A and B (and hence from C) to $A \star_C B$ are injective. One identifies A, B, and C with subgroups of $A \star_C B$, and one says that the amalgamated free product $A \star_C B$ is *nontrivial* if A and B are proper subgroups of $A \star_C B$.

Of course this comes up naturally in topology: if Z is, say, a connected bicollared hypersurface in an n-manifold M such that $M - Z$ has two connected components X and Y, and if the inclusion homomorphism from $\pi_1(Z)$ to $\pi_1(M)$ happens to be injective, then van Kampen's theorem exhibits $\pi_1(M)$ as the amalgamated free product $\pi_1(X) \star_{\pi_1(Z)} \pi_1(Y)$. (Here we should choose a base point in Z and take the homomorphisms i and j to be induced by inclusion.)

Generalizing the fact about free products which I talked about earlier in this section, one can show that any amalgamated free product $A \star_C B$ acts on a tree in such a way that the edge stabilizers are precisely the conjugates of C in $A \star_C B$, and the vertex stabilizers are precisely the conjugates of A and B. You can do this topologically by an argument very close to the one I gave for free products, or you can work it out algebraically from the normal form for elements of an amalgamated free product given in [40]. You can also find a proof in [58], about which I'll be saying more shortly. It's clear from the definitions that the action on the tree is nontrivial if and only if $A \star_C B$ is a nontrivial amalgamated free product.

If F is a separating, connected, essential surface in a closed, orientable, irreducible 3-manifold then $\pi_1(M)$ is a nontrivial free product with amalgamation. This amounts to saying that neither component of $M - F$ can carry $\pi_1(M)$, a fact which you can extract from the proof of 1.5.2 above, or deduce from the statement.

In the converse direction, if $\pi_1(M)$ is a nontrivial free product with amalgamation, then since $\pi_1(M)$ admits a nontrivial action an a tree, M must contain an essential surface. This is because $\pi_1(M)$ admits a nontrivial action on a tree, and you can apply the very first sentence of Section 2.5. However, the surface we get this way need not separate M. (Note that the dual surface to the action on the tree could fail to be connected.)

It's clear from these observations, together with the ones I made in Section 2.5, that a compact, orientable, irreducible 3-manifold is a Haken manifold if and only if *either* the first Betti number of M is strictly positive or $\pi_1(M)$ is a nontrivial free product with amalgamation. In [70], Waldhausen attributed this result to D.B.A. Epstein.

Feustel [30] showed that if $\pi_1(M) \cong A \star_C B$, where C is isomorphic to the fundamental group of a closed, orientable surface of positive genus, then there is a separating, closed, essential surface $F \subset M$ such that the inclusion homomorphisms map $\pi_1(F)$ onto C and map the fundamental groups of the components of $M - F$ onto A and B.

Of course there is no need to stop at free products with amalgamation. If Z is any bicollared hypersurface in an n-manifold M such that the inclusion homomorphism from $\pi_1(Z_i)$ is injective for each component Z_i of Z, one can use van Kampen's theorem to compute $\pi_1(M)$ from the fundamental groups of the components of Z and of $M - Z$. The appropriate structure was described elegantly in the work of Bass and Serre presented in [58], who introduced the notion of the "fundamental group of a graph of groups" as the relevant generalization of an amalgamated free product. As you would expect, the fundamental group of a graph of groups has a canonical action on a tree. What is more surprising is the converse: every action of an arbitrary group Γ on a tree arises in an essentially unique way from an isomorphism of Γ with the fundamental group of a graph of groups. There is also a topological approach to this theory; see [56].

You can think of the material in this chapter as being vaguely analogous to the Bass–Serre theory. If we start out, not with an abstract group but with the fundamental group of a (compact, irreducible, orientable) 3-manifold M, then we have seen how construct, from an action of $\pi_1(M)$ on a tree, not just a group-theoretical identification of $\pi_1(M)$ with the fundamental group of a graph of groups, but an essential surface in M itself, from which such an identification can be constructed. This construction is less canonical than the one given by the Bass–Serre theory, and the relationship between the dual surface and the action is less direct than the connection between a general group action and the associated graph of groups. It's nevertheless a useful construction. I've given a few hints in this section and the last one about why it's useful. To exploit it further one needs to combine it with ideas from algebra and geometry, which I'll be presenting in the next three sections.

3. The tree for SL_2

In the last two sections I illustrated how group actions on trees come up in 3-manifold theory. Another subject in which such actions come up naturally is the study of groups of 2×2 matrices with entries in a field: there is a natural way of constructing actions of such groups on trees, and this provides a beautiful and powerful way of analyzing the algebraic structure of these groups. In this section I will be giving a brief introduction to the ideas involved, from a purely algebraic point of view. In Section 5 I will explain the surprising interaction of these ideas from algebra with the topological theory of 3-manifolds.

The classic work on the tree for SL_2 is Serre's book [58]. The construction of the trees in question is a very special case of a construction due to Bruhat and Tits [13]. My aims here are to give a quick, self-contained account of enough of the material to allow you to read the rest of this chapter, and to inspire you to read [58] and perhaps [13].

3.1. *Valuations*

The starting point for this theory is the notion of a valuation. Valuations are objects that come up naturally in both number theory and complex analysis – so naturally, in fact, that anyone who has thought about the most elementary aspect of either subject has really worked with valuations, whether consciously or not. To illustrate the number-theoretic aspect of the idea, consider a prime number p. Given any integer $a \neq 0$, let us denote by $v(a) = v_p(a) \geqslant 0$ the exponent of p in the prime factorization of $|a|$. The fundamental theorem of arithmetic asserts that the surjective map $v : \mathbf{Z} - \{0\} \to \mathbf{N}$ (where $\mathbf{N}$ denotes the set of nonnegative integers) is well-defined. The reason the theorem is so powerful is that once we know v is well-defined, it follows immediately that it behaves well under addition and multiplication: we have

$$v(ab) = v(a) + v(b) \quad \text{for all } a, b \neq 0, \tag{3.1.1}$$

and

$$v(a+b) \geqslant \min\bigl(v(a), v(b)\bigr) \quad \text{for all } a, b \text{ such that } a, b \text{ and } a+b \text{ are all nonzero.} \tag{3.1.2}$$

(Inequality (3.1.2) just says that if a power of p divides both a and b then it divides $a+b$.)

Now it is an elementary exercise to show that if a map v of an integral domain R onto $\mathbf{N}$ satisfies (3.1.1) and (3.1.2), and if K denotes the field of fractions of R, then v extends uniquely to a map $\bar{v}$ of K onto $\mathbf{Z}$ satisfying the identities (3.1.1) and (3.1.2). (Any such $\bar{v}$ must satisfy $\bar{v}(a/b) = v(a) - v(b)$ for $a, b \in R - \{0\}$, and you can check that this formula gives a well-defined extension with the required properties. Surjectivity is clear.) Now if K is a field, a *valuation* of K is defined to be a surjection $v: K - \{0\} \to \mathbf{Z}$ that satisfies (3.1.1) and (3.1.2). (In the general theory of such things, these are called discrete, rank-1 valuations, but as they are the only kind of valuation I will be talking about for most of this chapter, I will just call them valuations for now. I will have occasion to mention more general valuations in Section 11.)

3.2. *The p-adic valuation*

So the map v_p of $\mathbf{Z}$ extends to a valuation of $\mathbf{Q}$. I'll denote the extension by v_p as well; it's called the p-adic valuation. If $a/b \in \mathbf{Q}$ is a fraction that has been written in lowest terms, and if p appears with exponent $r > 0$ in the factorization of a, we have $v_p(a/b) = r$. We have $v_p(a/b) = -r$ if p appears with exponent $r > 0$ in the factorization of b; and if p divides neither a nor b we have $v(a/b) = 0$.

As an illustration of how basic this is in number theory, consider the theorem, which was beyond the reach of the ancient Greeks even for $k = 2$, that the kth root of a positive integer n is either an integer or an irrational. The point is that if the kth root – let's call it α – is rational, then for every prime p we have $v_p(n) = kv_p(\alpha)$; thus $k | v_p(n)$ for every p, and it follows from the definition of the v_p that n is a kth power of an integer.

3.3. *Fields of meromorphic functions*

To illustrate the relevance of valuations to complex analysis, consider a meromorphic function f which is defined on an open set $u \subset \mathbf{C}$ and is not identically zero. For any point $z_0 \in u$ we can write $f(z)$, for z in some neighborhood of z_0, in a unique way as a Laurent series

$$f(z) = \sum_{n=r}^{\infty} a_n (z - z_0)^n,$$

where r is an integer, not necessarily positive, and $a_r \neq 0$. I'll call r the *order* of f at z_0. (If $r > 0$ we may say that f has a zero of order r, and if $r < 0$ that it has a pole of order $|r|$.) The meromorphic functions on u form a field K under pointwise addition and multiplication, and it's pretty clear that the function $v: K - \{0\} \to \mathbf{Z}$ that assigns to each meromorphic function its order at z_0 is a valuation.

When v is a valuation of a field K it's convenient to extend v to a function, still denoted v, defined on all of K and taking values in a set $\mathbf{Z} \cup \{\infty\}$, where ∞ is a new element that we adjoin to $\mathbf{Z}$, by setting $v(0) = \infty$. If we do this then the identities (3.1.1) and (3.1.2) hold for all $a, b \in K$, provided that we interpret $+$, $\geqslant$ and min in the obvious ways on the set $\mathbf{Z} \cup \{\infty\}$.

3.4. *The valuation ring*

A valuation v of a field K gives a good deal of nice structure. It's immediate from the definitions that the elements $x \in K$ such that $v(x) \geqslant 0$ form a sub-ring (with unity) of K. (I'm following the convention here according to which $v(0)$ is defined to be ∞ and therefore to be $\geqslant 0$.) This ring is called the *valuation ring* associated to v, and I'll denote it $\mathcal{O}_v$. Note that a nonzero element $x \in K$ satisfies $v(x) = 0$ if and only if x and x^{-1} are both in $\mathcal{O}_v$. Thus the elements $x \in K$ with $v(x) = 0$ comprise the (multiplicative) group of units $\mathcal{O}_v^*$ of the ring $\mathcal{O}_v$.

The ideals in the ring $\mathcal{O}_v$ are of a very simple form. To see this, let us fix an element $\pi \in K$ with $v(\pi) = 1$. Such an element is called a *uniformizer*. Let $\mathcal{I}$ be any ideal in $\mathcal{O}_v$, and let's set $n = \min_{x \in \mathcal{I}} v(x)$. If we fix an $x_0 \in \mathcal{I}$ with $v(x_0) = n$, then $\pi^n x_0^{-1} = n - n = 0$, so that $\pi^n \in x_0 \mathcal{O}_v^* \subset \mathcal{I}$. Conversely, for any $x \in \mathcal{I}$ we have $v(x/\pi^n) = v(x) - n \geqslant 0$, so that $\pi^n | x$ in $\mathcal{O}_v$. This shows that $\mathcal{I}$ is the principal ideal generated by π^n. So $\mathcal{O}_x$ is a principal ideal domain, and the only ideals are

$$(1) \supset (\pi) \supset \cdots \supset (\pi^n) \supset \cdots,$$

all of them linearly ordered by inclusion. In particular, $\mathcal{M}_v = (\pi)$, which consists of all elements $x \in K$ with $v(x) > 0$, is the unique maximal (proper) ideal of $\mathcal{O}_v$. Note that $\mathcal{O}_v - \mathcal{M}_v = \mathcal{O}_v^*$. Since $\mathcal{M}_v$ is a maximal ideal, the ring $k_v = \mathcal{O}_v / \mathcal{M}_v$ is a field, called the *residue field* of v.

Let's see what all this looks like in the examples of Sections 3.2 and 3.3. In the example of 3.2, the valuation ring is the ring $\mathbf{Z}_{(p)}$ consisting of all rational numbers which, when written in lowest terms, have denominators not divisible by the given prime p. In this example, p is itself a uniformizer, and the maximal ideal $p\mathbf{Z}_{(p)}$ of $\mathbf{Z}_{(p)}$ consists of all rational numbers which, when written in lowest terms, have numerators divisible by p. The residue field is easily seen to be isomorphic to $\mathbf{Z}/p\mathbf{Z}$; in fact, the usual homomorphism $\mathbf{Z} \to \mathbf{Z}/p\mathbf{Z}$ extends to a homomorphism $\mathbf{Z}_{(p)} \to \mathbf{Z}/p\mathbf{Z}$ with kernel $p\mathbf{Z}_{(p)}$.

In the example of Section 3.3, the valuation ring is the ring $\mathcal{O}_{z_0}$ consisting of all meromorphic functions on u which do not have poles at z_0. The function $z - z_0$ is a uniformizer, and the maximal ideal $\mathcal{M}_{z_0}$ of $\mathcal{O}_{z_0}$ consists of all meromorphic functions on u that have zeros at z_0. The homomorphism $f \mapsto f(z_0)$ from $\mathcal{O}_{z_0}$ to $\mathbf{C}$ is surjective, since it maps the field of constant functions isomorphically to $\mathbf{C}$, and its kernel is obviously $\mathcal{M}_{z_0}$. Hence the residue field k_v is canonically isomorphic to $\mathbf{C}$ in this example.

3.5. *The p-adics*

One way of thinking about a valuation v of a field K – and this is valuable general knowledge, although it won't be essential for the applications to 3-manifolds – is that it defines a "nonarchimedean absolute value" on K. Let's choose any real constant $c > 1$, and let's set $|x| = c^{-v(x)}$ for every $x \in K - \{0\}$ and $|0| = 0$. Then we have $|x| = 0$ if and only if $x = 0$, and the properties (3.1.1) and (3.1.2) of v translate into the identities

$$|ab| = |a| \cdot |b| \quad \text{and} \quad |a + b| \leqslant \max(|a|, |b|)$$

which hold for all $a, b \in K$. From this it is routine to deduce that K becomes a metric space if we define the distance between x and y to be $|x - y|$, and that the field operations are continuous in terms of the topology defined by this metric. The closed unit ball about 0 in K is obviously just the valuation ring $\mathcal{O}_v$. It's also routine to show that the field operations extend uniquely to the completion $\widehat{K}$ of the metric space K, which thereby itself becomes a field, and that v extends uniquely to a valuation $\hat{v}$ of $\widehat{K}$. It is not hard to show that the residue field of $\hat{v}$ is naturally isomorphic to the residue field of v.

The "p-adic distance" $d_p(x, y) = |x - y|_p$ defined by the valuation v_p of the field $\mathbf{Q}$ is an especially nice example. The definition of the distance depends on the choice of the constant c. For deep number-theoretic reasons it is customary to take the constant c to be p in this case. I'll use this choice of constant so as make my notation standard, but the choice of c does not affect anything I'll be talking about here. One nice feature of this example is that the unit ball $\mathcal{O}_v = \mathbf{Z}_{(p)}$ contains $\mathbf{Z}$ as a dense subset. To see this, note that if we're given any element of $\mathbf{Z}_{(p)}$, say a/b where $a, b \in \mathbf{Z}$ and p does not divide b, we have $p^n x + by = 1$ for some integers x and y, so that

$$\left|\frac{a}{b} - ay\right|_p = \left|\frac{axp^n}{b}\right| \leqslant p^{-n},$$

from which the assertion follows.

The field obtained by completing of $\mathbf{Q}$ with respect to the distance function d_p is called the field of p-adic numbers and is denoted $\mathbf{Q}_p$. The valuation ring of $\mathbf{Q}_p$, which is the closed unit ball about 0, is denoted $\mathbf{Z}_p$ and is called the ring of p-adic integers. Since $\mathbf{Z}$ is dense in $\mathbf{Z}_{(p)}$, it's easy to deduce that $\mathbf{Z}$ is dense in $\mathbf{Z}_p$ as well. So we can think of $\mathbf{Z}_p$ as the completion of $\mathbf{Z}$ with respect to the p-adic distance.

Now $\mathbf{Z}$ is a totally bounded metric space with respect to the p-adic distance. This is because for any integer $n > 0$ we can write $\mathbf{Z}$ as the union of p^n congruence classes modulo p^n, and each of these classes has diameter p^{-n}, so the completion $\mathbf{Z}_p$ of $\mathbf{Z}$ is compact. From this it's easy to deduce that $\mathbf{Q}_p$ is locally compact.

As a field with an absolute value, which is locally compact with respect to the topology defined by the absolute value, $\mathbf{Q}_p$ has many formal properties in common with the field $\mathbf{R}$ of real numbers. From a formal point of view it is interesting to ask questions about p-adic numbers that are analogous to familiar questions about real numbers. Actually such questions are of far more than formal interest, because of the role of the p-adic numbers in number theory. A famous example is the Hasse–Minkowski principle, which addresses the question of when a Diophantine equation in n variables

$$\sum a_{ij} x_i x_j = 0,$$

where a_{ij} are integers for $i, j \in \{1, \ldots, n\}$, has a nontrivial solution, i.e., whether there are integers $x_1, \ldots, x_n$, not all 0, that satisfy the equation. We can obviously replace both occurrences of the word "integers" here by "rational numbers", and we can assume without loss of generality that the matrix (a_{ij}) is symmetric; the question is then one about nontrivial zeros of a quadratic form in $\mathbf{Q}$. The Hasse–Minkowski principle says that such a form has a nontrivial zero in $\mathbf{Q}$ if (and only if) it has a nontrivial zero in $\mathbf{R}$ and also in

$\mathbf{Q}_p$ for every prime p. Saying the form has a nontrivial zero over $\mathbf{R}$ is the same as saying that it's indefinite – i.e., that it's neither positive definite nor negative definite – and this information can be read off from the signs of some minors of the matrix. The glorious part is that one can show, for example, that if $n \geqslant 5$ then the form has a nontrivial zero in $\mathbf{Q}_p$ for every p; so we get the elegant result that the above equation has a nontrivial integer solution whenever $n \geqslant 5$ and the (symmetrized) matrix of the form is indefinite.

The Hasse–Minkowski theorem is a special case of a "local-global" principle which says you can do certain kinds of things in $\mathbf{Q}$ if you can do them in $\mathbf{R}$ and in $\mathbf{Q}_p$ for every p. Of course the principle doesn't always work. Now you know as much about this as I do; to learn more, look at [57].

In Section 3.10 I'll give an actual proof of a p-adic analogue of a very familiar theorem involving real numbers.

3.6. *Defining the tree for* SL_2

Let K be any field. I'll show how to associate with any valuation v of K a tree $T = T_v$ on which $\mathrm{GL}(2, K)$ acts in a natural way. (As I've said, the tree is a special case of an object discovered by Bruhat and Tits. The description of it that I'll give is due to Serre.) Let's consider the standard 2-dimensional vector space $V = K^2$ over K. In particular we may regard V as a module over the valuation ring $\mathcal{O} = \mathcal{O}_v$. We define a *lattice* in V to be an $\mathcal{O}$-submodule of V which is finitely generated and spans V as a vector space over K. Since $\mathcal{O}$ is a principal ideal domain, any finitely generated $\mathcal{O}$-submodule of V is a free $\mathcal{O}$-module of some rank $\leqslant 2$. If the rank is < 2, the submodule cannot span V as a vector space. So any lattice is of rank 2: as far as their isomorphism type is concerned, all lattices look just like the standard lattice $\mathcal{O}^2 \subset K^2$.

In fact we can say slightly more. If Λ_0 and Λ_1 are lattices, and if $\{e_i, f_i\}$ is a basis of Λ_i as a free $\mathcal{O}$-module, then each $\{e_i, f_i\}$ is also a basis of V as a vector space over K. Hence there is a linear automorphism of V, which we can think of as an element A of $\mathrm{GL}(2, K)$, mapping $\{e_0, f_0\}$ onto $\{e_1, f_1\}$. In particular, A carries the lattice Λ_0 onto Λ_1.

When Λ_0 and Λ_1 are lattices, the element of $\mathrm{GL}(2, K)$ carrying Λ_0 onto Λ_1 is far from being unique, but it does define an invariant quantity according to the following result.

LEMMA 3.6.1. *Let Λ_0 and Λ_1 be lattices, and let A and B be two linear automorphisms of* $\mathrm{GL}(2, K)$ *that carry Λ_0 onto Λ_1. Then* $v(\det A) = v(\det B)$.

PROOF. Set $C = B^{-1}A$, so that $C(\Lambda_0) = \Lambda_0$. Hence if $\{e, f\}$ is a basis for Λ_0 as an $\mathcal{O}$-module, the matrix expressing C in terms of the basis $\{e, f\}$ has entries in $\mathcal{O}$, so that $\det C \in \mathcal{O}$. Since C^{-1} also leaves Λ_0 invariant, we have $(\det C)^{-1} \in \mathcal{O}$ as well. Thus $\det C$ is a unit in $\mathcal{O}$ and hence $v(\det C) = 0$. The conclusion of the lemma now follows from (3.1.1) and the multiplicativity of determinants. □

In view of this lemma we can associate an integer $\delta(\Lambda_0, \Lambda_1)$ with any ordered pair of lattices (Λ_0, Λ_1) by setting $\delta(\Lambda_0, \Lambda_1) = v(\det A)$, where A is an arbitrary linear automorphism of V mapping λ_0 onto Λ_1. From the multiplicativity of the determinant we see that

$$\delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_1, \Lambda_2) \tag{3.6.2}$$

for any lattices $\Lambda_0, \Lambda_1, \Lambda_2$. Since the identity has determinant 1, we also have

$$\delta(\Lambda, \Lambda) = 0 \tag{3.6.3}$$

for every lattice Λ.

If the lattices Λ_0 and Λ_1 satisfy $\Lambda_1 \subset \Lambda_0$, and if A is a linear transformation which maps Λ_0 onto Λ_1, then since $A(\Lambda_0) \subset \Lambda_0$, the argument used to prove Lemma 3.6.1 shows that in a suitable basis A has entries in $\mathcal{O}$ and therefore that $\det A \in \mathcal{O}$, i.e., $v(\det A) \geqslant 0$. So:

$$\text{If } \Lambda_1 \subset \Lambda_0 \text{ then } \delta(\Lambda_0, \Lambda_1) \geqslant 0. \tag{3.6.4}$$

Note also that if B is a linear transformation of V, then for any lattices Λ_0 and Λ_1 we have

$$\delta\big(B(\Lambda_0), B(\Lambda_1)\big) = \delta(\Lambda_0, \Lambda_1), \tag{3.6.5}$$

since if A is a linear transformation mapping Λ_0 onto Λ_1 then BAB^{-1} maps $B(\Lambda_0)$ onto $B(\Lambda_1)$.

We may think of $\delta(\cdot, \cdot)$ as an "algebraic distance" between lattices. What is more directly related to the construction of the tree T_v is a "geometric distance" between lattices, or more precisely between *homothety classes* of lattices. Two lattices $\Lambda, \Lambda' \subset V$ are said to be *(homothety)-equivalent,* or to represent the same homothety class, if there is a nonzero element α of K such that $\Lambda' = \alpha\Lambda$. You can see immediately that this really is an equivalence relation. It will turn out that our tree T_v is defined in such a way that its vertices are in bijective correspondence with homothety classes of lattices. The distance function that I will define will turn out to give the number of edges you have to follow to get from one vertex to another.

Because the homothety classes of lattices are going to be vertices, I will often use the letter s for a homothety class when I'm thinking of it as an object in its own right. (It's convenient here to follow Serre, who wrote in French: s means a vertex (*sommet*), whereas v is the valuation.) On the other hand, I will sometimes write $[\Lambda]$ for the homothety class of a lattice Λ that has already been named.

In order to define the geometric distance between homothety classes, we need two lemmas, of which the first is almost trivial.

LEMMA 3.6.6. *If Λ_0 and Λ_1 are lattices, then Λ_1 is equivalent to a lattice Λ_1' such that $\Lambda_1 \subset \Lambda_0$.*

PROOF. For $i = 0, 1$, let $\{e_i, f_i\}$ be a basis for Λ_i as an $\mathcal{O}$-module. Then $\{e_i, f_i\}$ is also a basis for V as a vector space, so we can write $e_1 = \alpha e_0 + \beta f_0$ for some $\alpha, \beta \in K$.

Since $e_1 \neq 0$ we have $m_0 = -\min(v(\alpha), v(\beta)) \in \mathbf{Z}$. For any $m \geqslant m_0$ we have $v(\pi^m\alpha) = m + v(\alpha) \geqslant 0$, so that $\pi^m\alpha \in \mathcal{O}$, and likewise $\pi^m\beta \in \mathcal{O}$. Hence $\pi^m e_1 \in \Lambda_0$ for $m \geqslant m_0$. Similarly, we see that $\pi^m f_1 \in \Lambda_0$ for m sufficiently large. So we can find an m for which $\pi^m e_1$ and $\pi^m f_1$ both belong to Λ_0; hence $\pi^m \Lambda_1$, which is equivalent to Λ_1, is contained in Λ_0. □

RAPPEL 3.6.7. In the proof of the next lemma I'll be using a basic result on finitely generated modules over a principal ideal domain: *if L_0 is a free module of finite rank over a p.i.d. R, and L_1 is a submodule of L_0, then there exist a basis $\{e_1, \ldots, e_n\}$ for L_0 and elements $\alpha_1, \ldots, \alpha_n$ of R such that L_1 is generated by $\alpha_1 e_1, \ldots, \alpha_n e_n$.* (This result underlies one proof of the structure theorem for finitely generated modules over R: any finitely generated submodule M can obviously be written in the form L_0/L_1, where L_0 is a finitely generated free module and L_1 is a submodule of L_0; the above result then shows that M is a finite direct sum of cyclic modules.)

LEMMA 3.6.8. *If Λ_0 and Λ_1 are lattices, then there is a unique lattice Λ_1' equivalent to Λ_1 such that $\Lambda_1 \subset \Lambda_0$ and Λ_0/Λ_1 is isomorphic as an $\mathcal{O}$-module to $\mathcal{O}/\beta\mathcal{O}$ for some nonzero element β of $\mathcal{O}$.*

PROOF. By Lemma 3.6.6, we may assume that Λ_1 is already contained in Λ_0. By the result I mentioned before the proof, in Rappel 3.6.7, Λ_0 has a basis $\{e, f\}$ such that Λ_1 is generated by $\{\alpha e, \gamma f\}$ for some $\alpha, \gamma \in \mathcal{O}$. After possibly reversing the roles of e and f we may assume that $v(\alpha) \leqslant v(\gamma)$; setting $\beta = \gamma\alpha^{-1}$ we conclude that $v(\beta) \geqslant 0$, so that $\beta \in \mathcal{O}$. The lattice $\Lambda_1' = \alpha^{-1}\Lambda_1$, which is equivalent to Λ, has the basis $\{e, \beta f\}$. It follows that $\Lambda_1' \subset \Lambda_0$ and that Λ_0/Λ_1' is isomorphic as an $\mathcal{O}$-module to $\mathcal{O}/\beta\mathcal{O}$. This proves the existence assertion.

Now suppose there is a second lattice Λ_1'' equivalent to Λ_1 such that $\Lambda_1'' \subset \Lambda_0$ and such that Λ_0/Λ_1'' is isomorphic as an $\mathcal{O}$-module to $O/\delta\mathcal{O}$ for some $\delta \in \mathcal{O}$. We may write $\Lambda_1'' = \zeta\Lambda_1'$ for some $\zeta \in F$. Since Λ_1' contains the basis element e of $\mathcal{O}$, we have $\zeta e \in \Lambda_1'' \subset \Lambda_0$, which implies that $\zeta \in \mathcal{O}$. Since Λ_1'' is generated by ζe and $\zeta\beta f$, we have $\Lambda_0/\Lambda_1'' \cong \mathcal{O}/\zeta\mathcal{O} \oplus \mathcal{O}/\zeta\beta\mathcal{O}$. But the uniqueness part of the structure theorem for modules over a p.i.d. implies that $\mathcal{O}/\zeta\mathcal{O} \oplus \mathcal{O}/\zeta\beta\mathcal{O}$ can't be cyclic unless ζ is a unit in $\mathcal{O}$, in which case $\Lambda_1'' = \Lambda_1'$. This proves the uniqueness assertion, and completes the proof of Lemma 3.6.8. □

Given lattices Λ_1 and Λ_0, I will say that Λ_1 is *snugly embedded in* Λ_0 if Λ_0 and Λ_1 are related in the way described in the conclusion of Lemma 3.6.8, that is, if $\Lambda_1 \subset \Lambda_0$ and Λ_0/Λ_1 is a cyclic $\mathcal{O}$-module. Now if s_0 and s_1 are homothety classes of lattices, I'll define the "geometric distance" $d(s_0, s_1)$ to be the integer $\delta(\Lambda_0, \Lambda_1)$, where the Λ_i are representatives of the s_i such that Λ_1 is snugly embedded in Λ_0. According to Lemma 3.6.8, such representatives Λ_0 and Λ_1 exist, and Λ_1 is uniquely determined once Λ_0 has been chosen. To show that $d(s_0, s_1)$ is independent of the choice of Λ_0, note that if Λ_0 and Λ_0' are representatives of s_0, so that $\Lambda_0' = \alpha\Lambda_0$ for some nonzero element α of F, and if Λ_1 is a representative of s_1 that's snugly embedded in Λ_0, then $\Lambda_1' = \alpha\Lambda_1$ represents s_1 and is snugly embedded in Λ_0', and by applying (3.6.5) to the linear transformation $x \mapsto \alpha x$ we find that

$$\delta(\Lambda_0', \Lambda_1') = \delta(\alpha\Lambda_0, \alpha\Lambda_1) = \delta(\Lambda_0, \Lambda_1).$$

It follows from (3.6.4) that $d(s_0, s_1) \geqslant 0$ for any two homothety classes of lattices s_0 and s_1.

To understand the definition of d better, let's consider an arbitrary lattice Λ_0 and a lattice Λ_1 that's snugly embedded in Λ_0. From the proof of Lemma 3.6.8 (or, more precisely, the existence part of the proof and the uniqueness part of the statement) we see that Λ_0 has a basis $\{e, f\}$ such that e and βf generate Λ_1 for some $\beta \in F - \{0\}$. The linear transformation of V whose matrix in the basis $\{e, f\}$ is $\begin{pmatrix} 1 & 0 \\ 0 & \beta \end{pmatrix}$ has determinant β and maps Λ_0 onto Λ_1. Hence

$$d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1) = v(\beta). \tag{3.6.9}$$

It's worth noticing that if π is a uniformizer in $\mathcal{O}_v$ then in the above discussion we may take β to be a nonnegative power of π, since every nonzero element of $\mathcal{O}_v$ is a nonnegative power of π with a unit. If $\beta = \pi^n$ then $d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1) = n$.

It's useful to generalize the description of the distance that I just gave. If Λ_0 and Λ_1 are lattices with $\Lambda_1 \subset \Lambda_0$, then by Rappel 3.6.7 Λ_0 and Λ_1 have bases of the form $\{e, f\}$ and $\{\alpha e, \gamma f\}$ for some $\alpha, \gamma \in F - \{0\}$. Here, using the matrix $\begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}$, we find that

$$\delta(\Lambda_1, \Lambda_0) = v(\alpha\gamma) = v(\alpha) + v(\gamma).$$

This is used in the proof of the following lemma, which is the main step in the proof that d is a distance function.

LEMMA 3.6.10. *If $\Lambda_0 \supset \Lambda_1$ are lattices, we have*

$$d([\Lambda_0], [\Lambda_1]) \leqslant \delta(\Lambda_0, \Lambda_1)$$

and

$$d([\Lambda_0], [\Lambda_1]) \equiv \delta(\Lambda_0, \Lambda_1) \pmod 2.$$

Furthermore, we have $d([\Lambda_0], [\Lambda_1]) = \delta(\Lambda_0, \Lambda_1)$ if and only if Λ_1 is snugly embedded in Λ_0.

PROOF. This is a lot like the existence part of the proof of Lemma 3.6.8. We fix bases $\{e, f\}$ and $\{\alpha e, \gamma f\}$ for Λ_0 and Λ_1, where $\alpha, \gamma \in \mathcal{O}$. As in the proof of 3.6.8, we may assume that $v(\alpha) \leqslant v(\gamma)$; and as in that proof it follows that $\beta = \gamma\alpha^{-1} \in \mathcal{O}$, and that $\Lambda_1' = \alpha^{-1}\Lambda_1$, which is equivalent to Λ_1, has the basis $\{e, \beta f\}$ and is therefore snugly embedded in Λ_0. By the remarks before the statement of the lemma we're proving now, we find that

$d([\Lambda_0],[\Lambda_1]) = \delta(\Lambda_0, \Lambda_1') = v(\beta)$, and that $\delta(\Lambda_1, \Lambda_0) = v(\alpha) + v(\gamma) = v(\beta) + 2v(\alpha)$, so that

$$\delta(\Lambda_0, \Lambda_1) = d\big([\Lambda_0], [\lambda_1]\big) + 2v(\gamma),$$

from which the first two assertions follow. If $\delta(\Lambda_0, \Lambda_1) = d([\Lambda_0], [\lambda_1])$ then $v(\gamma) = 0$; thus $\gamma \in \mathcal{O}^*$, and it follows that $\{e, \gamma f\}$ is a basis for Λ_1, hence that Λ_1 is snugly embedded in Λ_0. The converse follows from the definition of d. □

Let's denote by $T^{(0)}$ the set of all homothety classes of lattices, so that d is a nonnegative integer-valued function on $T^{(0)} \times T^{(0)}$.

LEMMA 3.6.11. *$(T^{(0)}, d)$ is a metric space.*

PROOF. Since any lattice is obviously snugly embedded in itself, it follows from (3.6.3) that $d(s, s) = 0$ for any s. Conversely, if $d(s_0, s_1) = 0$, and if we represent the s_i by lattices Λ_i where Λ_1 is snugly embedded in Λ_0, then Λ_0 and Λ_1 have bases $\{e, f\}$ and $\{e, \beta f\}$ for some $\beta \in \mathcal{O}$ with $v(\beta) = d([\Lambda_0], [\Lambda_1]) = 0$; hence $\beta \in \mathcal{O}^*$, from which it follows that $\Lambda_1 = \Lambda_0$ and hence $s_0 = s_1$.

To prove symmetry we consider two arbitrary elements s_0, s_1 of $T^{(0)}$, which we represent by lattices Λ_0 and Λ_1, where Λ_1 is snugly embedded in Λ_0. Again we choose a basis $\{e, f\}$ of Λ_0 such that e and $f' = \beta f$ form a basis of Λ_1. So $d(s_0, s_1) = \delta(\Lambda_0, \Lambda_1) = v(\beta)$. On the other hand, $\Lambda_0' = \beta\Lambda_0$ also represents s_0, and it has the basis $\{f', \beta e\}$; hence Λ_0' is snugly embedded in Λ_1, and $d(s_1, s_0) = \delta(\Lambda_1, \Lambda_0') = v(\beta)$.

To prove the triangle inequality we consider arbitrary elements v_0, v_1, v_2 of $T^{(0)}$. If Λ_0 is any lattice representing s_0 then successive applications of Lemma 3.6.8 give lattices Λ_1 and Λ_2 representing s_1 and s_2, with Λ_{i+1} snugly embedded in Λ_i for $i = 0, 1$. Now $\Lambda_2 \subset \Lambda_0$, and by Lemma 3.6.10 and (3.6.2) we have

$$d(s_0, s_2) \leqslant \delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_0, \Lambda_2) = d(s_0, s_1) + d(s_1, s_2),$$

where the last equality follows directly from the definition of d. □

Before I can move on to the definition of the tree T_v, I need to establish a couple of other properties of the metric space $T^{(0)}$. Like the last result, they are applications of Lemma 3.6.10.

LEMMA 3.6.12. *For any $s_0, s_1, s_2 \in T^{(0)}$ we have*

$$d(s_0, s_2) \equiv d(s_0, s_1) + d(s_1, s_2) \pmod 2.$$

PROOF. By successive applications of Lemma 3.6.6, we can represent the s_i by lattices Λ_i with $\Lambda_2 \subset \Lambda_1 \subset \Lambda_0$. By (3.6.2) and the second assertion of Lemma 3.6.10, we find that

$$\begin{aligned} d(s_0, s_2) &\equiv \delta(\Lambda_0, \Lambda_2) = \delta(\Lambda_0, \Lambda_1) + \delta(\Lambda_1, \Lambda_2) \\ &\equiv d(s_0, s_1) + d(s_1, s_2) \pmod 2. \end{aligned}$$

□

LEMMA 3.6.13. *Let s_0 and s_1 be elements of $T^{(0)}$, set $n = d(s_0, s_1)$, and let p and q be nonnegative integers with $p + q = n$. Then there is a unique element s of $T^{(0)}$ such that $d(s_0, s) = p$ and $d(s, s_1) = q$.*

PROOF. Let's represent the s_i by lattices Λ_i, with Λ_1 snugly embedded in Λ_0. There are bases $\{e, f\}$ and $\{e, \beta f\}$ of Λ_0 and Λ_1, and we may take $\beta = \pi^n$ where π is a uniformizer and $n = d(s_0, s_1)$. If we now define Λ to be the lattice generated by e and $\pi^p f$, it is clear that Λ is snugly embedded in Λ_0 and that Λ_1 is snugly embedded in Λ_1. By (3.6.9), the element $s = [\Lambda]$ of $T^{(0)}$ has the required properties.

Now suppose that some $s' \in T^{(0)}$ satisfies $d(s_0, s) = p$ and $d(s, s_1) = q$. We wish to prove that $s' = s$. By successive applications of Lemma 3.6.8, there exist a representative Λ' of s which is snugly embedded in Λ_0, and a representative Λ'_1 of s_1 which is snugly embedded in Λ. Using the definition of d and (3.6.2), we find that

$$\begin{aligned} d\big([\Lambda_0], [\Lambda'_1]\big) &= n = p + q = d(s_0, s) + d(s', s_1) \\ &= \delta\big(\Lambda_0, \Lambda'\big) + \delta\big(\Lambda, \Lambda'_1\big) = \delta\big(\Lambda_0, \Lambda'_1\big), \end{aligned}$$

which by Lemma 3.6.10 implies that Λ'_1 is snugly embedded in Λ_0. It now follows from the uniqueness assertion of Lemma 3.6.8 that $\Lambda'_1 = \Lambda_1$. Thus we have $\Lambda_1 \subset \Lambda \subset \Lambda_0$.

The lattices that contain Λ_1 and are contained in Λ_0 are in bijective correspondence with the submodules of $\Lambda_0/\Lambda_1 \cong \mathcal{O}/\pi^n\mathcal{O}$. Since every ideal in $\mathcal{O}$ is generated by a power of π, the only submodules of Λ_0/Λ_1 are those generated by π^k for $0 \leqslant k \leqslant n$. Hence every lattice that contains Λ_1 and is contained in Λ_0 is generated by e and $\pi^k\beta$ for some $k \leqslant n$. In particular Λ' has this form for some k. But then by (3.6.9) we have $k = d(s_0, s) = p$, so $\Lambda' = \Lambda$ and hence $s' = s$. Lemma 3.6.13 is now proved. □

Let's define an abstract simplicial 1-complex $T = T_v$ as follows. The set of vertices of T is $T^{(0)}$. A 1-simplex is an unordered pair (s, s') of vertices such that $d(s, s') = 1$. I'll use the same name T (or T_v) to refer to the geometric realization of this complex; and as is usual in such situations, it will be either be clear from the context which I mean, or it won't matter.

THEOREM 3.6.14. *The 1-complex T is 1-connected, i.e., it is a tree.*

PROOF. Let s and s' be any two elements of $T^{(0)}$. Set $k = d(s, s')$. By successive applications of the existence assertion of Lemma 3.6.13, we find elements $s = s_0, s_1, \ldots, s_k = s'$ of $T^{(0)}$ such that $d(s_{i-1}, s_i) = 1$ for $i = 1, \ldots, k$. This defines an edge path between the vertices s, s' of T and shows that T is connected.

To show that T is simply connected we must show that for every *reduced* edge path $s_0, \ldots, s_n$ of length $n > 0$ in T we have $s_n \neq s_0$. To say that the edge path is reduced means that in addition to having $d(s_{i-1}, s_i) = 1$ for $i = 1, \ldots, n$, we have $s_{i-1} \neq s_{i+1}$ whenever $0 < i < n$. What I'll prove, by induction on n, is that if $s_0, \ldots, s_n$ is any reduced edge path of length $n > 0$ then $d(s_0, s_n) = n$. For $n = 1$ this is trivial. Now suppose that $s_0, \ldots, s_{n+1}$ is a reduced edge path of length $n + 1$, where $n > 0$, and assume that the assertion is true for shorter paths, so that $d(s_0, s_n) = n$ and $d(s_0, s_{n-1}) = n - 1$. Since $d(s_n, s_{n+1}) = 1$, the triangle inequality gives $n - 1 \leqslant d(s_0, s_{n+1}) \leqslant n + 1$. By Lemma 3.6.12, we have

$$d(s_0, s_{n+1}) \equiv d(s_0, s_n) + d(s_n, s_{n+1}) = n + 1 \pmod 2,$$

so we can't have $d(s_0, s_{n+1}) = n$. It remains to rule out the possibility that $d(s_0, s_{n+1}) = n - 1$. Assume that this does hold. Then we have

$$d(s_0, s_{n-1}) = d(s_0, s_{n+1}) = n - 1,$$
$$d(s_{n-1}, s_n) = d(s_{n+1}, s_n) = 1,$$

and

$$d(s_0, s_n) = n.$$

Invoking the uniqueness assertion of Lemma 3.6.13, taking $p = n - 1$, $q = 1$, and letting s_n play the role of s_1, we conclude that $s_{n+1} = s_{n-1}$. But this contradicts the assumption that $s_0, \ldots, s_{n+1}$ is a reduced edge path. The proof of Theorem 3.6.14 is now complete. □

3.7. *The action*

Now it's very easy to bring $\mathrm{GL}(2, K)$ into the picture. We can think of an element B of $\mathrm{GL}(2, K)$ as a linear automorphism of $V = K^2$. As such, B maps any lattice onto a lattice, and it obviously maps equivalent lattices to equivalent lattices. So there is a natural action of $\mathrm{GL}(2, K)$ on the set $T^{(0)}$ of homothety classes of lattices. It's also clear that if Λ_1 is snugly embedded in Λ_0 then $B \cdot \Lambda_1$ is snugly embedded in $B \cdot \Lambda_0$, and by (3.6.5) we have

$$d([B \cdot \Lambda_0, B \cdot \Lambda_1]) = \delta(B \cdot \Lambda_0, B \cdot \Lambda_1) = \delta(\Lambda_0, \Lambda_1) = d([\Lambda_0, \Lambda_1]),$$

so that $\mathrm{GL}(2, K)$ acts by isometries on $T^{(0)}$. In particular, each element of $\mathrm{GL}(2, K)$ carries 1-simplices onto 1-simplices, so that we have a natural action of $\mathrm{GL}(2, K)$ on the tree T.

In this chapter I will mostly be using the action of $\mathrm{SL}(2, K)$ on T that comes from restricting the action of $\mathrm{GL}(2, K)$. There are a couple of points to be made about this action of $\mathrm{SL}(2, K)$. First of all, $\mathrm{SL}(2, K)$ acts on T without inversions. This is because if s is any vertex of T, and Λ is any lattice representing s, then for any $B \in \mathrm{SL}(2, K)$ we have $d(s, B \cdot s) \equiv \delta(\Lambda, B \cdot \Lambda) = 0 \pmod 2$ by Lemma 3.6.10 and the definition of δ; in particular we always have $d(s, B \cdot s) \neq 1$, so B can't act as an inversion.

The second point to be made about the action of $\mathrm{SL}(2, K)$ is that the stabilizers of vertices have a very simple description. Let s be any vertex, let B be an element of the stabilizer $\mathrm{SL}(2, K)_s$, and let Λ be a lattice representing s. Then $B \cdot s$ is homothety-equivalent to s, so $B \cdot s = \alpha s$ for some $\alpha \in K - \{0\}$. Since $\delta(\cdot, \cdot)$ is well-defined, we have $2v(\alpha) = \delta(\Lambda, B \cdot \Lambda) = v(\det B) = v(1) = 0$. Hence α is a unit in $\mathcal{O}$, so that $B \cdot \Lambda = \Lambda$. Conversely, if $B \in \mathrm{SL}(2, K)$ leaves Λ invariant, it is obvious that $B \cdot s = s$. Thus $\mathrm{SL}(2, K)_s$ is the stabilizer of Λ. Now the stabilizer of the standard lattice $\mathcal{O}^2$ is the group $\mathrm{SL}(2, \mathcal{O})$

(consisting of all 2×2 matrices of determinant 1 with entries in $\mathcal{O}$). If A is an element of $\mathrm{GL}(2, K)$ such that $A \cdot \mathcal{O}^2 = \Lambda$, we have $\mathrm{SL}(2, K)_s = \mathrm{SL}(2, \mathcal{O})^A$, where exponentiation denotes conjugation. What I've shown is that the stabilizers in $\mathrm{SL}(2, K)$ of the vertices of T are just the conjugates of $\mathrm{SL}(2, \mathcal{O})$ in $\mathrm{GL}(2, K)$.

3.8. *Getting to know the tree, I: The link of a vertex*

A good starting point for understanding what the tree $T = T_v$ looks like is describing the link of a vertex. Let $s_0 = [\Lambda_0] \in T^{(0)}$ be given. The link of s_0 consists of all elements s of $T^{(0)}$ such that $d(s_0, s) = 1$. Any such s is represented by a unique lattice Λ which is snugly embedded in Λ_0; thus if π denotes a uniformizer in $\mathcal{O}_v$, the lattice Λ_0 has a basis $\{e, f\}$ (depending on s) such that $\{e, \pi f\}$ is a basis of Λ_0. So we have

$$\pi \Lambda_0 \subset \Lambda \subset \Lambda_0. \tag{3.8.1}$$

Now the lattices Λ that satisfy (3.8.1) are in bijective correspondence with submodules of the quotient module $V = \Lambda_0 / \pi \Lambda_0$, which we can think of as a 2-dimensional vector space over the residue field $k = k_v = \mathcal{O}_v / \pi \mathcal{O}_v$ of v. If Λ is in fact generated by e and πf for some basis $\{e, f\}$ of λ_0, then the corresponding subspace of V is clearly 1-dimensional; and the converse is easy to deduce from Rappel 3.6.7. So we get a canonical bijection between vertices in the link of s_0 and 1-dimensional subspaces of V.

The set of 1-dimensional subspaces of a 2-dimensional vector space V over k is, by definition, a 1*-dimensional projective space*, or *projective line*, over k. I'll return to projective spaces in Section 5.2. For now let me just observe that if V has a basis we can identify the corresponding projective line with the disjoint union of k with a single element denoted ∞: the subspace spanned by the vector whose coordinates in the basis are a and b is identified with $a/b \in k$ if $b \neq 0$, and with ∞ if $b = 0$.

In the case where k is a finite field with q elements – for example when v is the p-adic valuation of $\mathbf{Q}$ or $\mathbf{Q}_p$ and $q = p$ – the set $k \cup \{\infty\}$ has $q + 1$ elements, and hence each vertex of T has valence $q + 1$. A good exercise in understanding the tree is to take $q = 2$, so that $k \cong \mathbf{Z}/2\mathbf{Z}$ and T is a trivalent tree, and, starting with an arbitrary vertex, to describe some nearby vertices. Suppose we denote by $[e, f]$ the vertex represented by the lattice generated by a given basis $\{e, f\}$ of F^2. If we write a given vertex in the form $[e, f]$, the vertices in its link are $[e, 2f]$, $[2e, f]$ and $[2e, e + f]$. (The latter vertex could equally well have been written as $[e + f, 2f]$.) Now we can find the three vertices in the link of, say, $[2e, e + f]$ by substituting $2e$ and $e + f$ for e and f in the expression for the vertices in the link of $[e, f]$; doing this directly gives $[2e, 2e + 2f]$, $[4e, f]$ and $[4e, 3e + f]$. However, $[2e, 2e + 2f]$ is simply the original vertex $[e, f]$ under a different name, which makes sense because we already know that $[e, f]$ and $[2e, e + f]$ are joined by an edge. So the two new vertices in the link of $[2e, e + f]$ are $[4e, f]$ and $[4e, 3e + f]$, which you may prefer to rename $[4e, -e + f]$. You can continue in this way and see various interesting new vertices appear at small distances from $[e, f]$.

Since the action of $\mathrm{SL}(2, K)$ on T is simplicial, it restricts to an action of the stabilizer $\mathrm{SL}(2, K)_s$ of any vertex s on the link of s. Up to equivalence, what we are looking at here is an action of $\mathrm{SL}(2, \mathcal{O})$ on the standard 1-dimensional projective space $k\mathrm{P}^1$ over the residue field k. Once we've said that, it's pretty clear what this action should be (again up to equivalence): the quotient homomorphism $\mathcal{O} \to k$ gives rise to a natural homomorphism $q : \mathrm{SL}(2, \mathcal{O}) \to \mathrm{SL}(2, k)$, and $\mathrm{SL}(2, k)$ acts in a natural way on $k\mathrm{P}^1$ because a linear transformation of the vector space k^2 permutes the 1-dimensional subspaces of k^2. (If you identify $k\mathrm{P}^1$ with $k \cup \{\infty\}$ then $\mathrm{SL}(2, K)$ acts on $k\mathrm{P}^1$ by linear fractional transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \to \frac{az+b}{cz+d} \quad \text{for } z \in k \cup \{\infty\}$$

as you can check by an easy calculation.) The obvious action of $\mathrm{SL}(2, \mathcal{O})$ on $k\mathrm{P}^1$ is obtained by pulling back this action of $\mathrm{SL}(2, K)$ via q. As is usual in such situations, proving that this natural action really is equivalent to the one obtained by restricting the action of $\mathrm{SL}(2, K)$ on T is just an exercise in keeping track of the definitions.

One consequence of this description of the action of $\mathrm{SL}(2, K)_s$ on the link of s is a description of the stabilizer in $\mathrm{SL}(2, K)$ of an edge of T. If e is an edge with endpoints s_0 and s_1, we can think of the stabilizer $\mathrm{SL}(2, K)_e$ of e as the stabilizer of s_1 within the group $\mathrm{SL}(2, K)_{s_0}$. But under the standard action of $\mathrm{SL}(2, K)$ on $k\mathrm{P}^1$, the stabilizer of a point of $k\mathrm{P}^1$ is conjugate to the group Δ of upper triangular matrices in $\mathrm{SL}(2, K)$. It follows that $\mathrm{SL}(2, K)_e$ is conjugate in $\mathrm{GL}(2, K)$ to $q^{-1}(\Delta) \subset \mathrm{SL}(2, \mathcal{O})$. The latter group can be described directly as consisting of all matrices of the form $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $c \in \pi\mathcal{O}$ and $a, b, d \in \mathcal{O}$.

This in turn has a neat consequence concerning the commutator subgroup $[\mathrm{SL}(2, K)_e, \mathrm{SL}(2, K)_e]$ of an edge stabilizer $\mathrm{SL}(2, K)_e$, which will come up a couple of times in the applications that I'll talk about later in this chapter. We have $q([\mathrm{SL}(2, K)_e, \mathrm{SL}(2, K)_e]) \subset [\Delta, \Delta]$, and the group $[\Delta, \Delta]$ is made up entirely of upper triangular matrices over k that have 1's on the diagonal. In particular, if e is an edge of T_v we have

$$\operatorname{trace} A \equiv 2 \pmod 2 \quad \text{for every } A \in \big[\mathrm{SL}(2, K)_e, \mathrm{SL}(2, K)_e\big]. \tag{3.8.2}$$

3.9. *Getting to know the tree, II: Axes*

Now that we have a good picture of the stabilizer of a vertex and how it acts on the link of the vertex, let's ask the opposite question: what can we say about the action of an element A of $\mathrm{SL}(2, K)$ that fixes no vertex of T? For simplicity, let's assume that A is diagonalizable over K, so that for some basis $\{e, f\}$ of K^2 we have $A(e) = \alpha e$ and $A(f) = \alpha^{-1} f$. The assumption that A fixes no vertex says that α and α^{-1} do not both belong to $\mathcal{O}$; by symmetry we may assume $\alpha^{-1} \notin \mathcal{O}$, so that $l = v(\alpha) > 0$. Using the notation of Section 3.8, let us set $s_n = [e, \pi^n f] \in T^{(0)}$ for every $n \in \mathbf{Z}$. We have $d(s_m, s_n) = |m - n|$ for all $m, n \in \mathbf{Z}$. It follows that there is an edge a_n joining s_n to s_{n+1} for each n, and that the s_n and a_n form a subcomplex L of T which, up to simplicial isomorphism, looks like the real line triangulated so that the integers are the vertices. The definition of the action of $\mathrm{SL}(2, K)$ on T implies that $A \cdot s_n = s_{n+2l}$ for every $n \in \mathbf{Z}$. (In fact, A maps the lattice generated by e and $\pi^n f$ onto the one generated by $\pi^{-l}e$ and $\pi^{n+l}f$.) So the "simplicial line" L is invariant under A, and A acts on L as a translation.

It's a neat combinatorial exercise to prove that if T is any tree and γ is any simplicial automorphism of T which is not an inversion and fixes no vertex of T, then there is a unique subcomplex L of T which is simplicially isomorphic to a line and is invariant under γ; furthermore, γ always acts on L as a translation. You will find this worked out in [58], or in a more general version in [45]. The line L is called the *axis* of γ. So what I've done here is to describe the axis in T of a diagonalizable element of $SL(2, K)$ (when it exists, i.e., when the element has no fixed point in T).

3.10. *Application: Ihara's theorem*

To illustrate what can be done with the tree for SL_2, I will give Serre's elegant proof of a result due to Ihara which is the p-adic analogue of a simple fact about discrete subgroups of $SL(2, \mathbf{R})$. Suppose that $\Gamma \subset SL(2, \mathbf{R})$ is a discrete, torsion-free group. Then Γ maps injectively to $PSL(2, \mathbf{R})$, the group of orientation-preserving isometries of the hyperbolic plane $\mathbf{H}^2$; the image is still discrete and of course torsion-free. From this one deduces that the action of Γ on $\mathbf{H}^2$ is free. This is because the stabilizer Γ_z of each point $z \in H^2$ is a discrete subgroup of the compact group $SL(2, \mathbf{R})_z \cong SO_2$, and is therefore finite, hence trivial since Γ is torsion-free. The discreteness of Γ also implies that Γ acts properly discontinuously on $\mathbf{H}^2$, so the quotient $\mathbf{H}^2/\Gamma$ is an orientable hyperbolic surface F having $\mathbf{H}^2$ as its universal covering space and Γ as its group of deck transformations. Hence $\Gamma \cong \pi_1(F)$. This shows that every discrete, torsion-free subgroup of $SL(2, \mathbf{R})$ either is a free group or is isomorphic to the fundamental group of a closed orientable surface of genus $\geqslant 2$.

Ihara's theorem says that in the p-adic world the corresponding result is even simpler: every discrete, torsion-free subgroup of $SL(2, \mathbf{Q}_p)$ is free! (Here the term "discrete" is to be interpreted in terms of the topology on $SL(2, \mathbf{Q}_p)$ defined in the obvious way from the metric topology of $\mathbf{Q}_p$.) The proof closely parallels the one for the real case, but it uses the tree T for $SL(2, \mathbf{Q}_p)$ in place of $\mathbf{H}^2$. First we show Γ acts freely on the set of vertices of T: this is formally identical to the proof in the real case that Γ acts freely on $\mathbf{H}^2$, once one knows that the stabilizer $SL(2, \mathbf{Q}_p)_s$ is compact for every vertex; but by what we saw in Section 3.7, $SL(2, \mathbf{Q}_p)_s$ is conjugate in $GL(2, \mathbf{Q}_p)$ to $SL(2, \mathbf{Z}_p)$; and as we saw in Section 3.5 that $\mathbf{Z}_p$ is compact, the group $SL(2, \mathbf{Z}_p)$ is clearly compact as well. Now that we know that Γ acts freely on the vertices of T, it follows that it acts freely on the whole geometric simplicial complex T because, according to 3.7, there are no inversions. So the quotient T/Γ is a graph G having T as its universal covering space and Γ as its group of deck transformations. (Note that since the action is simplicial this time, proper discontinuity is not even an issue.) Hence $\Gamma \cong \pi_1(G)$, and it follows that Γ is free.

4. Varieties of representations and varieties of characters

I talked in Section 1.6 about the nearly canonical representation of the fundamental group of a finite-volume hyperbolic 3-manifold M in $PSL(2, \mathbf{C})$ or $SL(2, \mathbf{C})$. It turns out that when M has cusps, this representation can be "deformed" through infinite families of inequivalent representations which can be studied with the techniques of algebraic geometry. The punch line, later in the chapter, is going to be that deforming representations "off to infinity" produces actions of $\pi_1(M)$ on trees, which are defined using the construction of Section 3, and which in turn can be used to define incompressible surfaces in M using the constructions of Section 2.

In this section I'll try to provide a rough introduction to the needed foundational ideas from algebraic geometry, as well as presenting the more specialized material involving representations and hyperbolic manifolds. Although I won't be able to make this section as self-contained as Section 3, I'll try to give a hint of what the required material is about, and to provide references, where necessary, to sources where proofs are presented in an accessible form.

I've decided to present this theory from the point of view taken in [20,21,18,22], involving SL(2, **C**)-representations. It has been shown, for example in [9,15], that stronger information can sometimes be obtained using PSL(2, **C**)-representations; but this requires taking a less elementary point of view, and you may have your hands full already.

4.1. *The variety of representations*

Let Γ be any finitely generated group. We are interested in studying the set $R(\Gamma)$ of all representations of Γ in SL(2, **C**). Suppose we fix a finite system of generators of Γ, say $(g_1, \ldots, g_n)$. Then a representation $\rho : \Gamma \to \mathrm{SL}(2, \mathbf{C})$ is uniquely determined by specifying the n-tuple $(\rho(g_1), \ldots, \rho(g_n))$. Here each $\rho(g_i)$ is a matrix

$$\begin{pmatrix} w_i & g_i \\ y_i & z_i \end{pmatrix} \in \mathrm{SL}(2, \mathbf{C}),$$

so we may think of ρ as being determined by the $4n$-tuple $(w_1, x_1, y_1, z_1, \ldots, w_n, x_n, y_n, z_n)$ of complex numbers. This gives a bijective correspondence $\rho \leftrightarrow (\rho(g_1), \ldots, \rho(g_n))$ between $R(\Gamma)$ and some subset of the complex affine space $\mathbf{C}^{4n}$. It will be useful to think of $R(\Gamma)$ as being identified with this subset of $\mathbf{C}^{4n}$ via this correspondence. (Of course this identification depends on choosing a system of generators $(g_1, \ldots, g_n)$ for Γ. I'll return to this issue in Section 4.3.)

Now suppose that $(r_j)_{j \in J}$ is a system of defining relators for Γ; here the index set J may be finite or infinite, and each r_j is a word in the generators $g_1, \ldots, g_n$. If $X_1, \ldots, X_n$ are 2×2 matrices, we denote by $r_j(X_1, \ldots, X_n)$ the matrix that's obtained by substituting X_i for g_i in the word r_j for $i = 1, \ldots, n$. Then a $4n$-tuple $(w_1, \ldots, z_n)$ belongs to the set $R(\Gamma)$ if and only if we have

$$w_i z_i - x_i y_i = 1 \tag{4.1.1}$$

for $i = 1, \ldots, n$, and

$$r_j\left(\begin{pmatrix} w_1 & x_1 \\ y_1 & z_1 \end{pmatrix}, \ldots, \begin{pmatrix} w_n & x_n \\ y_n & z_n \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.1.2}$$

for each $j \in J$. For each i Equation (4.1.1) is a polynomial equation in the coordinates of $\mathbf{C}^{4n}$. For each j we can rewrite the matrix equation (4.1.2) as a system of four polynomial equations in the coordinates. To do this, we first rewrite each occurrence of an inverse matrix

$$\begin{pmatrix} w_i & x_i \\ y_i & z_i \end{pmatrix}^{-1} \quad \text{as} \quad \begin{pmatrix} z_i & -x_i \\ -y_i & w_i \end{pmatrix}$$

(which is equal to $\left(\begin{smallmatrix} w_i & x_i \\ y_i & z_i \end{smallmatrix}\right)^{-1}$ in the presence of Equations (4.1.1)). Then we multiply out the left-hand side of (4.1.2) and set each of the four matrix entries of the resulting product equal to the corresponding matrix entry on the right-hand side.

This shows that the set $R(\Gamma) \subset \mathbf{C}^{4n}$ is the solution set to some system of polynomial equations in the coordinates of $\mathbf{C}^{4n}$. In general, a subset of an affine space $\mathbf{C}^N$ is called a *(complex affine) algebraic set* if it's the set of zeros of some system of polynomial equations in the coordinates. The set of defining equations that we have exhibited for $R(\Gamma)$ may be infinite, if Γ is not finitely presented; however, one of the first things that one proves in algebraic geometry – a corollary to the Hilbert basis theorem – is that any subset of $\mathbf{C}^N$ which is defined by a possibly infinite system of polynomial equations can actually be defined by some finite subsystem.

4.2. *A little algebraic geometry*

This is a good place to review a few basic concepts and results concerning algebraic sets in an affine space $\mathbf{C}^N$, where N is a natural number. (The proofs of these facts can be found in any introductory text on algebraic geometry. One book that I have found congenial is [48].) An algebraic set is said to be *reducible* if it can be expressed as the union of two proper algebraic subsets. Irreducible affine algebraic sets are often called *affine varieties.*

Another corollary of the Hilbert basis theorem states that any algebraic set $V \subset \mathbf{C}^N$ is a finite union of irreducible algebraic sets $V_1 \cup \cdots \cup V_k$. Once this has been established, it's obvious that we can choose the V_i so that $V_i \not\subset V_j$ whenever i and j are distinct indices $\leqslant k$. With this restriction, it isn't hard to show that the decomposition $V = V_1 \cup \cdots \cup V_k$ is unique apart from the order of the terms. The V_i are called the *irreducible components* of V. Unlike the connected components of a topological space, the irreducible components of an algebraic set V are not necessarily disjoint from one another. (The algebraic subset of $\mathbf{C}^2$ defined by the equation $zw = 0$, where z and w denote the coordinates, has the coordinate axes $z = 0$ and $w = 0$ as its irreducible components.)

If $V \subset \mathbf{C}^N$ is any algebraic set, the *coordinate ring* $\mathbf{C}[V]$ of V is defined, most concretely, to be the ring of all functions on V which are restrictions of functions on $\mathbf{C}^N$ defined by polynomials in the coordinates. Since $\mathbf{C}[V]$ contains the constant functions we may think of it as an algebra over $\mathbf{C}$, and as such it is generated by the restrictions to V of the coordinate functions on $\mathbf{C}^N$; in particular it is a finitely generated $\mathbf{C}$-algebra. If V

is irreducible, it's a simple exercise in using the definitions to show that $\mathbf{C}[V]$ is an integral domain. In this case one denotes by $\mathbf{C}(V)$ the field of fractions of the integral domain $\mathbf{C}[V]$.

If V is an affine variety, any element of $\mathbf{C}(V)$ may be written in the form f/g, where $f, g \in \mathbf{C}(V)$ and $g \neq 0$ (i.e., the function g does not vanish identically on V). The points of V where g takes the value 0 form a proper algebraic subset of V. It is a general fact that any proper algebraic subset of an affine variety V is made up of irreducible components having lower dimension than V, and has a dense complement in V. Thus g is nonzero on an open dense subset U of V. The given element of $\mathbf{C}(V)$ defines a function on U whose value at a point $x = (z_1, \ldots, z_N)$ is $f(z_1, \ldots, z_N)/g(z_1, \ldots, z_N)$.

This leads to an alternative description of the elements of $\mathbf{C}(V)$ as equivalence classes of genuine functions. Each of the functions in question is required to have a domain which is the complement a proper algebraic subset of V, and on this domain it is required to be defined by rational functions in the coordinates of the affine space containing V. Two such functions are equivalent if they agree on the intersection of their domains.

For this reason, the elements of $\mathbf{C}(V)$ are called *rational functions* on V, and $\mathbf{C}(V)$ is referred to as the *function field* of V.

A *polynomial map* between algebraic sets $V \subset \mathbf{C}^M$ and $W \subset \mathbf{C}^N$ is a map $F : V \to W$ which is defined by polynomials in the ambient coordinates. More precisely, F is a polynomial map if there are elements $f_1, \ldots, f_N$ of $\mathbf{C}[V]$ such that $F(x) = (f_1(x), \ldots, f_N(x))$ for every $x \in V$. A priori, of course, if $f_1, \ldots, f_N$ are elements of $\mathbf{C}[V]$ then $(f_1(x), \ldots, f_N(x))$ is only a point of $\mathbf{C}^N$; to say that it lies in W for every $x \in V$ says that the f_i satisfy certain algebraic relations.

It's often useful to think of affine algebraic sets as forming a category, with polynomial maps playing the role of morphisms. In particular we have a natural notion of *isomorphism* of affine algebraic sets. From this point of view, the coordinate ring behaves like (you should excuse the expression) a contravariant functor: if $F : V \to W$ is a polynomial map, then for every $g \in \mathbf{C}[W]$, the function $g \circ F$ belongs to $\mathbf{C}[V]$. (We're just composing polynomials to get another polynomial.) The map $g \mapsto g \circ F$ is a homomorphism of $\mathbf{C}$-algebras from $\mathbf{C}[W]$ to $\mathbf{C}[V]$. It's obvious that if F is surjective – or more generally if it maps V onto a dense subset of W – then the associated homomorphism $\mathbf{C}[W] \to \mathbf{C}[V]$ is injective. So if V and W are irreducible, there is an induced homomorphism (necessarily injective!) of fields of fractions, from $\mathbf{C}(W)$ to $\mathbf{C}(V)$. So when we have fixed a polynomial map of V onto W, where V and W are irreducible, we can think of the field $\mathbf{C}(V)$ as an extension of $\mathbf{C}(W)$.

4.3. *More on varieties of representations*

Let's return to the study of the set of representations $R(\Gamma)$ of a finitely generated group Γ. I have pointed out that fixing a system of generators $(g_1, \ldots, g_n)$ of Γ defines a bijection, say η, of $R(\Gamma)$ onto an affine algebraic set in $\mathbf{C}^{4n}$. By definition we have $\eta(\rho) = (\rho(g_1), \ldots, \rho(g_n))$. Now suppose that $(h_1, \ldots, h_m)$ is a second system of generators of Γ, and let $\theta : \rho \mapsto (\rho(h_1), \ldots, \rho(h_m))$ denote the corresponding bijection to an algebraic set in $\mathbf{C}^{4m}$. The composition $\theta \circ \eta^{-1}$ is a bijection between algebraic sets. Let's write $h_i = W_i(g_1, \ldots, g_n)$ for $i = 1, \ldots, m$, where each W_i is a word in n letters; and let's identify a point $(w_1, x_1, y_n, z_n, \ldots, w_n, x_n, y_n, z_n)$ of $\mathbf{C}^{4n}$ with an n-tuple

$$\left(\begin{pmatrix} w_1 & g_1 \\ y_1 & z_1 \end{pmatrix}, \ldots, \begin{pmatrix} w_n & g_n \\ y_n & z_n \end{pmatrix}\right)$$

of 2×2 matrices, and likewise for $\mathbf{C}^{4m}$. Then the composite bijection $\theta \circ \eta^{-1}$ maps an n-tuple of matrices $(X_1, \ldots, X_n)$ to the m-tuple

$$\big(W_i(X_1, \ldots, X_n)\big)_{i=1}^{m}.$$

Because matrix multiplication and inversion involve only multiplying and adding entries and changing signs, it follows that $\theta \circ \eta^{-1}$ is a polynomial map. The same argument shows that $\eta \circ \theta^{-1}$ is a polynomial map. So the natural bijection between the two algebraic sets incarnating $R(\Gamma)$ is an isomorphism of algebraic sets; this means that the structure of an algebraic set that we have given to $R(\Gamma)$ is really a completely natural one.

I'll ordinarily be identifying $R(\Gamma)$ with an actual algebraic set in an affine space by fixing some set of generators, and the remark I just made says that nothing algebro-geometric about $R(\Gamma)$ really depends on the set of generators. I find it very reassuring to know that, although I don't know if I'll actually use it anywhere in this chapter. On the other hand, there is another remark I need to make about the algebraic set $R(\Gamma)$ which is absolutely fundamental for the mathematics I'll be talking about.

Suppose R_0 is an algebraic subset of $R(\Gamma)$, for example $R(\Gamma)$ itself or an irreducible component. Suppose we fix an element $\gamma \in \Gamma$. Then every $\rho \in R_0$ defines a matrix $\rho(\gamma) \in \mathrm{SL}(2, \mathbf{C})$. Since we are thinking of γ as being fixed, the entries of $\rho(\gamma)$ are determined by the element ρ of R_0; so we can write

$$\rho(\gamma) = \begin{pmatrix} a_\gamma(\rho) & b_\gamma(\rho) \\ c_\gamma(\rho) & d_\gamma(\rho) \end{pmatrix}, \tag{4.3.1}$$

where a, b, c, d are complex-valued functions on R_0 determined by the element γ. Since each $\rho \in R_0$ is a representation, we have $\rho(\gamma\delta) = \rho(\gamma)\rho(\delta)$, i.e.,

$$\begin{pmatrix} a_{\gamma\delta}(\rho) & b_{\gamma\delta}(\rho) \\ c_{\gamma\delta}(\rho) & d_{\gamma\delta}(\rho) \end{pmatrix} = \begin{pmatrix} a_\gamma(\rho) & b_\gamma(\rho) \\ c_\gamma(\rho) & d_\gamma(\rho) \end{pmatrix} \begin{pmatrix} a_\delta(\rho) & b_\delta(\rho) \\ c_\delta(\rho) & d_\delta(\rho) \end{pmatrix} \tag{4.3.2}$$

for all $\rho \in R_0$ and $\gamma, \delta \in \Gamma$.

Now if as usual we think of $R(\Gamma)$ as being a concrete set in an affine space by fixing a set of generators for Γ, and if the element γ happens to be a generator, then it is immediate from (4.3.1) that $a_\gamma, b_\gamma, c_\gamma, d_\gamma$ are just the restrictions to R_0 of the four coordinate functions corresponding to that generator. So they belong to the coordinate ring $\mathbf{C}[R_0]$. For an arbitrary element γ, if we write out γ as a word in the generators and repeatedly apply (4.3.2), then again because of the polynomial nature of matrix multiplication and inversion, we conclude that $a_\gamma, b_\gamma, c_\gamma, d_\gamma \in \mathbf{C}[R_0]$ for every $\gamma \in \Gamma$. Notice also that for any $\gamma \in \Gamma$ and any $\rho \in R_0$ we have $a_\gamma(\rho)d_\gamma(\rho) - b_\gamma(\rho)c_\gamma(\rho) = \det \rho(\gamma) = 1$; so $a_\gamma d_\gamma - b_\gamma c_\gamma = 1$

for every $\gamma \in \Gamma$. This means that the matrix $\begin{pmatrix} a_\gamma & b_\gamma \\ c_\gamma & d_\gamma \end{pmatrix}$ is an element of $\mathrm{SL}(2, \mathbf{C}[r_0])$ for every γ. Finally, it follows from (4.3.2) that the map $\mathcal{P}: \Gamma \to \mathrm{SL}(2, \mathbf{C}[r_0])$ defined by

$$\mathcal{P}(\gamma) = \begin{pmatrix} a_\gamma & b_\gamma \\ c_\gamma & d_\gamma \end{pmatrix} \tag{4.3.3}$$

is a homomorphism, i.e., a representation of Γ in $\mathrm{SL}(2, \mathbf{C}[r_0])$. In particular, if R_0 is irreducible and if K denotes the field $\mathbf{C}(R_0)$, we may regard $\mathcal{P}$ as a representation in $\mathrm{SL}(2, K)$.

In [20] Culler and I named $\mathcal{P}$ the *tautological representation*, and topologists working in this area have generally used this term, although a very similar object is sometimes referred to by algebraists as a *universal representation*.

A central theme in this chapter, which first appeared in my joint paper [20] with Culler, is that the representation $\mathcal{P}$ can be used, via the theory that I presented in Section 3, to define actions of Γ on trees. But before I can explain how this works, I need to introduce a little more machinery.

4.4. *Varieties of characters*

From the beginning of the chapter I've been stressing the theme that one is interested in representations primarily *up to equivalence*. However, the elements of the set $R(\Gamma)$ are representations in $\mathrm{SL}(2, \mathbf{C})$, not equivalence classes of representations, and for some purposes this is a defect. In this subsection I'll show how to parametrize the *characters* of representations of a finitely generated group Γ by points of an affine algebraic set $X(\Gamma)$, much as the representations themselves are parametrized in the way I described in Section 4.1. According to Proposition 1.1.1, characters almost classify representations up to equivalence, so the points of $X(\Gamma)$ are very nearly in bijective correspondence with equivalence classes of representations.

We can think of the group $\mathrm{SL}(2, \mathbf{C})$ as acting on $R(\Gamma)$ by conjugation: for any $A \in \mathrm{SL}(2, \mathbf{C})$ and for any representation $\rho \in R(\Gamma)$ we can define $A \cdot \rho = i_A \circ \rho$, where i_A is the inner automorphism $X \mapsto AXA^{-1}$. The equivalence classes of representations are the orbits of this action. Furthermore, the action is algebraic in the sense that the map $(A, \rho) \mapsto A \cdot \rho$ is a polynomial map from $\mathrm{SL}(2, \mathbf{C}) \times R(\Gamma)$ to $R(\Gamma)$, as you can easily check by arguments like the ones I gave in Section 4.1.

The fancy-delancey point of view about the algebraic set $X(\Gamma)$ is that it is the quotient of $R(\Gamma)$ by the action of $\mathrm{SL}(2, \mathbf{C})$, *in the category of algebraic sets*. (Because inequivalent representations can sometimes have the same character, it is *not* the quotient in the category of "sets, period".) There is a general theory of quotients under group actions in algebraic geometry, called geometric invariant theory, which certainly subsumes the material I'll be covering in this section. The point of view I'll be presenting here is closely modeled on the point of view that Culler and I used in [20]. I will be adopting this point of view partly because I don't know geometric invariant theory, and partly because I want to show how elementary the material is. At one point I will refer you to [20] for a result that we proved using the "Burnside Lemma", but that is also quite elementary algebra.

To begin with, let's define a function $I_\gamma : R(\Gamma) \to \mathbf{C}$ for each $\gamma \in \Gamma$, by setting $I_\gamma(\rho) = \operatorname{trace} \rho(\gamma)$ for every representation $\rho \in R(\Gamma)$. Using the notation of Section 4.3, with $R_0 = R(\Gamma)$, we deduce from (4.3.1) that $I_\gamma(\rho) = a_\gamma(\rho) + d_\gamma(\rho)$; comparing this with (4.3.3), we conclude that

$$I_\gamma = \operatorname{trace} \mathcal{P}(\gamma) \tag{4.4.1}$$

for every $\gamma \in \Gamma$. In particular, I_γ is an element of the coordinate ring $\mathbf{C}[R(\Gamma)]$ for every $\gamma \in \Gamma$.

I'll define the *trace ring* $T(\Gamma)$ to be the sub-ring of $\mathbf{C}[R(\Gamma)]$ generated by all the functions I_γ for $\gamma \in \Gamma$. (By definition the elements of $T(\Gamma)$ are functions that can be expressed as integer polynomials in the I_γ.) The following elementary result provides a finite set of generators for $T(\Gamma)$ as a ring.

PROPOSITION 4.4.2. *Suppose that a group Γ is generated by elements $\gamma_1, \dots, \gamma_n$. Then the trace ring $T(\Gamma)$ is generated by the elements I_V, where V ranges over all elements of the form $\gamma_{i_1} \cdots \gamma_{i_k}$ with $1 \leqslant k \leqslant n$ and $1 \leqslant i_1 < \cdots < i_k \leqslant n$. (Note that this set of generators of $T(\Gamma)$ has $2^n - 1$ elements.)*

PROOF. The proof is based on the identity

$$\operatorname{trace} AB + \operatorname{trace} AB^{-1} = (\operatorname{trace} A)(\operatorname{trace} B), \tag{4.4.3}$$

which holds for all $A, B \in \mathrm{SL}(2, \mathbf{C})$. This identity has a beautiful proof which I learned from Troels Jorgensen. The characteristic polynomial of A is $X^2 - (\operatorname{trace} A)X + 1$, so the Cayley–Hamilton theorem says that $A^2 - (\operatorname{trace} A)A + I = 0$, i.e., $A + A^{-1} = (\operatorname{trace} A)I$. Now multiply both sides on the right by B and take traces to get (4.4.3).

We can interpret (4.4.3) in terms of the functions I_γ as saying that for any $\gamma, \gamma' \in \Gamma$ we have

$$I_{\gamma'\gamma} + I_{\gamma'\gamma^{-1}} = I_\gamma I_{\gamma'}; \tag{4.4.4}$$

in fact, if we evaluate both sides of (4.4.4) at a point ρ of $R(\Gamma)$ we get (4.4.3) with $A = \rho(\gamma')$ and $B = \rho(\gamma)$.

Let's denote by T_0 the sub-ring of $T(\Gamma)$ generated by elements of the special form described in the statement of the proposition. Using (4.4.4) we can prove by induction on the length of a word W in the generators $\gamma_1, \dots, \gamma_n$ that $I_W \in T_0$. This will give the conclusion. You can think of the induction as starting at length 0, where the assertion is trivial because I_1 (where 1 means the identity element of Γ) is the constant function 2. Now consider a word W of length $n > 0$, and assume the assertion is true for words of length $< n$. We can assume W is a reduced, since otherwise we can replace it by a shorter word representing the same element of Γ.

Suppose that W' is a word obtained from W by inverting a single letter somewhere in W: that is, W has the form $X\gamma_i^\varepsilon Y$ as a word, for some $i \leqslant n$ and $\varepsilon = \pm 1$, and $W' = X\gamma_i^{-\varepsilon} Y$. (When I say that W has the form $X\gamma_i^\varepsilon Y$ *as a word*, the juxtaposition of X, γ_i^ε, Y

represents concatenation of words and not merely multiplication in the group. In particular, $n = \text{length}\, W = \text{length}\, X + \text{length}\, Y + 1$. Note that the word W' need not be reduced.) I claim that $I_W \in T_0$ if and only if $I_{W'} \in T_0$. This is because we can rewrite I_W and $I_{W'}$ as $I_{YX\gamma_i^\varepsilon}$ and $I_{YX\gamma_i^{-\varepsilon}}$ in view of the familiar identity trace $AB =$ trace BA, and then by (4.4.4) we find that

$$I_W + I_{W'} = I_{YX} I_{\gamma^\varepsilon}.$$

Since $I_{XY} \in T_0$ by the induction hypothesis, and since $I_{\gamma^\varepsilon} = I_\gamma$ is one of the generators of T_0, the claim follows.

Next I claim that we can interchange two successive letters in W without affecting the membership of I_W in T_0; that is, if $W = X\gamma_i^\varepsilon \gamma_j^\zeta Y$ as a word, and $W' = X\gamma_j^\zeta \gamma_i^\varepsilon Y$, then I_W belongs to T_0 if and only if $I_{W'}$ does. This is because the same argument used to prove my last claim shows that $I_W \in T_0$ if and only if $I_{X(\gamma_i^\varepsilon \gamma_j^\zeta)^{-1}Y} = I_{X\gamma_j^{-\zeta}\gamma_i^{-\varepsilon}Y}$ belongs to T_0, so this claim now follows from the last one.

Now by repeatedly interchanging successive letters we can replace W by a word which either fails to be reduced or has the form $\gamma_1^{k_1} \cdots \gamma_n^{k_n}$ for some $k_1, \ldots, k_n \in \mathbf{Z}$. If we assume W to have the latter form then after possibly inverting certain letters we can assume the k_i to be nonnegative. If some k_i is $\geqslant 2$, we can invert a single letter and get a nonreduced word. So we can assume W already has the form $\gamma_1^{k_1} \cdots \gamma_n^{k_n}$ with each k_i equal to either 0 or 1. But in this case I_W is by definition a generator of T_0. This proves the proposition. □

Now, given a finitely generated group Γ, let's fix a set of generators $\gamma_1, \ldots, \gamma_n$ for Γ. Setting $N = 2^n - 1$, let's index the words of the form $\gamma_{i_1} \cdots \gamma_{i_k}$, with $1 \leqslant k \leqslant n$ and $1 \leqslant i_1 < \cdots < i_k \leqslant n$, in some order as $V_1, \ldots, V_N$. We define a map $t : R(\Gamma) \to \mathbf{C}^N$ by $t(\rho) = (I_{V_1}(\rho), \ldots, I_{V_n}(\rho))$. If two points ρ, ρ' in $R(\Gamma)$ have the same image under t, i.e., if $I_{V_i}(\rho) = I_{V_i}(\rho')$, then it follows from Proposition 4.4.2 that $I_\gamma(\rho) = I_\gamma(\rho')$ for every $\gamma \in \Gamma$. By the definition of the I_γ this means that trace $\rho(\gamma) =$ trace $\rho'(\gamma)$ for every γ, i.e., that ρ and ρ' have the same character. Conversely, if ρ and ρ' have the same character then in particular $t(\rho) = t(\rho')$. So the points of $t(R(\Gamma))$ are in natural bijective correspondence with the characters of representations of Γ in $\mathrm{SL}(2, \mathbf{C})$, and the map t sends each representation to the point corresponding to its character. From now on I will identify $t(R(\Gamma))$ with the set of characters of representations of Γ, just as I identified the set of representations itself with a subset of $\mathbf{C}^{4n}$ in Section 4.1.

Whereas it was essentially obvious that $R(\Gamma) \subset \mathbf{C}^{4n}$ was an algebraic set, the corresponding fact for characters requires more work. I will refer you to [20] for a proof, using the "Burnside Lemma", that $t(R(\Gamma)) \subset \mathbf{C}^N$ is an algebraic set. For a still more elementary proof of this, see [32]. From this point I will denote the algebraic set $t(R(\Gamma))$ by $X(\Gamma)$.

Since t maps $R(\Gamma)$ onto $X(\Gamma)$, we have a natural injective homomorphism $J : \mathbf{C}[X(\Gamma)] \to \mathbf{C}[R(\Gamma)]$ by Section 4.2. The algebra $\mathbf{C}[R(\Gamma)]$ is generated by the restrictions of the coordinate functions in $\mathbf{C}^N$. The homomorphism J carries the ith coordinate function to its composition with t, which by the definition of t is just I_{V_i}. So the ring $J(\mathbf{C}[R(\Gamma)])$ is generated by the I_{V_i}. According to Proposition 4.4.2 it follows that $J(\mathbf{C}[R(\Gamma)])$ coincides with the sub-algebra $\mathbf{C}T[R(\Gamma)]$ generated by the functions I_γ for $\gamma \in \Gamma$. In particular, each I_γ is in the image of J, that is, it is obtained from a polynomial function on $X(\Gamma)$ by composition with t.

I'll generally just identify $\mathbf{C}[X(\Gamma)]$ with its image under J. This means that each function f on $X(\Gamma)$ is identified with $J(f) = f \circ t$. As a special case, the function on $X(\Gamma)$ from which I_γ is obtained by composition will also be denoted I_γ. In this language we can say that the functions I_γ generate the algebra $\mathbf{C}[X(\Gamma)]$.

4.5. *The irreducible component of a discrete faithful character*

Having introduced the formalism of the character variety, I can now be precise about the ideas I was waving my hands about in the introduction to this section. Let N be an orientable hyperbolic 3-manifold of finite volume. According to Proposition 1.6.1, a (discrete, faithful) representation $\rho_0 : \pi_1(N) \to \mathrm{PSL}(2, \mathbf{C})$ associated to the hyperbolic structure of N admits a lift $\tilde{\rho}_0 : \pi(N) \to \mathrm{SL}(2, \mathbf{C})$. Thinking of $\tilde{\rho}_0$ as a point of $R(\pi_1(N))$, we get an associated point $\chi_0 = t(\tilde{\rho}_0) \in X(\pi_1(N))$. It should be clear at this point that when I talked in the introduction to the section about ways of "deforming $\tilde{\rho}_0$ through inequivalent representations", I was referring to the study of the irreducible component(s) of $X(\pi_1(N))$ containing χ_0.

When N is closed, everything that can be said about the subject is contained in two theorems [68,69] due to Weil. Whenever N has finite volume, whether or not it is closed, the main result of [69], the "local rigidity theorem", implies that any discrete faithful representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N))$ is equivalent to $\tilde{\rho}_0$ as a representation in $\mathrm{SL}(2, \mathbf{C})$. This is forerunner of the "strong rigidity theorem" later proved by Mostow, and can be easily deduced from it; Mostow's theorem asserts in this context that any two discrete faithful representations of $\pi_1(N)$ in $\mathrm{Isom}(\mathbf{H}^3)$ are equivalent. (See Bonahon's chapter in this Handbook [6].)

In the case where N is closed, the results of [68] imply that any representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N)$ is still discrete and faithful. So Weil's results, taken together, say in this particular context that an entire neighborhood of $\tilde{\rho}_0$ in $R(\pi_1(N))$ is contained in a single equivalence class of representations, and hence maps to a point in $X(\pi_1(M))$. This suggests that the image χ_0 of $\tilde{\rho}_0$ should be an isolated point in $X(\pi_1(N))$, and this can in fact be deduced from Weil's results with a little fiddling.

An (irreducible) affine variety is always connected according to [48, Corollary 4.16], so the isolated point χ must constitute a 0-dimensional irreducible component of $X(\pi_1(N))$, which is of course the only irreducible component containing χ_0. In the informal language of the introduction to the section, $\tilde{\rho}_0$ cannot be deformed through inequivalent representations when N is closed.

The correct generalization of this to the case of a finite-volume manifold which may have cusps is essentially due to Thurston. In one version, it states that if N is an orientable, finite-volume, hyperbolic 3-manifold with n cusps, and if $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ is defined as above, then any irreducible component X_0 of $X(\pi_1(N))$ containing $\chi_0 = t(\rho_0)$ has dimension n. (This is not the strongest known version; more about this at the end of this subsection.)

Thurston's proof of this is divided into two parts. In the first step, which is elementary, ingenious, and essentially algebraic, one shows that any component of $X(\pi_1(N))$ which contains χ_0 must have dimension at least n. The idea is to write down a presentation of $\pi_1(N)$ from which one can deduce that $X = X(\pi_1(N))$, which is realized concretely as an algebraic set in some affine space $\mathbf{C}^r$, can be defined "in a neighborhood of χ_0" by $r - n$ equations. (More precisely, this means that there is a polynomial map $f : \mathbf{C}^r \to \mathbf{C}^{r-n}$ such that $f^{-1}(0) \cap U = X \cap U$ for some neighborhood U of χ_0 in $\mathbf{C}^r$.) This implies the required lower bound on dimension by virtue of general facts about algebraic sets. The only facts used in describing X locally by the right kind of system of equations are that N is homeomorphic to the interior of a compact manifold M whose boundary consists of tori $B_0, \ldots, B_{n-1}$, and that the representation $\tilde{\rho}_0$ is irreducible and maps each of the subgroups $\mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ isomorphically to a group of parabolic elements in $\mathrm{SL}(2, \mathbf{C})$.

You will find an account of this part of the argument in [20] (proof of Proposition 3.2.1).

I want to say a little more about the second part of Thurston's argument because it gives additional information which will be important in this chapter. This part uses hyperbolic geometry, and the key step is an adaptation of the main theorem of [68] to the case of a finite-volume hyperbolic manifold with cusps. In terms of the notation that I just introduced, the relevant result is that if a representation $\tilde{\rho} : \pi_1(N) \to \mathrm{SL}(2, \mathbf{C})$ is sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N))$, *and* if ρ shares with $\tilde{\rho}_0$ the property that it maps each of the groups $P_i = \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ (defined up to conjugacy, and often called *peripheral subgroups*) onto a group of parabolic elements in $\mathrm{SL}(2, \mathbf{C})$, then ρ is still discrete and faithful. By [69], ρ is then equivalent to ρ_0, if it is close enough to it.

Like the corresponding statement in the closed case, this one is easily translated into a statement about the character variety. Let X^* denote the algebraic subset of $X(\pi_1(N))$ obtained by adjoining the additional equations $I_\gamma^2 = 4$, for all conjugacy classes represented by elements of the subgroups P_i, to the defining equations for $X(\pi_1(N))$. So X^* consists of all characters of representations that send all the peripheral subgroups onto groups of parabolic elements. The translation of the adapted version of Weil's theorem is that χ_0 is an isolated point of X^*.

Now suppose that for each $i \leqslant n$ we fix a nontrivial element γ_i of P_i. (You should think of the elements $\gamma_1, \ldots, \gamma_n$ as being defined up to conjugacy, as the peripheral subgroups are.) It is elementary to see that in a neighborhood of χ_0, the algebraic set X^* is defined by adding the equations $I_{\gamma_i}^2 = 4$, for $i = 0, \ldots, n$, to the defining equations for X. (The main point, at least intuitively, is that if $\rho \in R(\pi_1(N))$ is a representation sufficiently close to $\tilde{\rho}_0$ in $R(\pi_1(N))$, such that $I_{\gamma_1}, \ldots, I_{\gamma_n}$ vanish at $t(\rho)$, then each $\rho(\gamma_i)$ is a nontrivial element of $\mathrm{SL}(2, \mathbf{C})$ with trace ± 2. Any other element γ of P_i commutes with $\rho(\gamma_i)$ and must therefore also have trace ± 2, so that I_γ vanishes at $t(\rho)$. Again, translating this into the required statement requires a bit of fiddling.)

Now it's a basic fact about complex affine varieties that if a variety X has dimension d then any irreducible component of a subset of X defined by n additional equations has dimension at least $d - n$. (See [59, p. 60, Theorem 7].) If a component X_0 containing χ_0 had dimension $d > n$, then since X_* is defined in the neighborhood of χ_0 by the n additional equations $I_{\gamma_i} = 0$, any component of X_* containing χ_0 would have dimension $d - n > 0$. This is a contradiction since χ_0 is isolated in X^*.

For details, see [21].

This argument gives more information than the statement about dimension that I gave at the outset. Let's summarize what it shows:

THEOREM 4.5.1. *Let N be an orientable hyperbolic 3-manifold of finite volume, and n denote the number of its cusps. Let $\tilde{\rho}_0 : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ be a lift of a representation $\rho_0 : \pi_1(N) \to \mathrm{PSL}(2, \mathbf{C})$ associated to the hyperbolic structure of N, set $\chi_0 = t(\tilde{\rho}_0) \in X(\pi_1(N))$, and let X_0 be an irreducible component of $X(\pi_1(N))$ containing χ_0. Then $\dim X_0 = n$. Furthermore, if $B_0, \ldots, B_{n-1}$ are the boundary components of a compact core of N, if γ_i is an element whose conjugacy class is carried by B_i for $i = 0, \ldots, n-1$, then χ_0 is an isolated point of the algebraic subset*

$$X^* = \{\chi \in X_0 \colon I_{\gamma_1}^2 = \cdots = I_{\gamma_n}^2 = 4\}$$

of X_0.

This is the theorem I will be quoting in later sections. I should point out that stronger statements are generally believed to be true: that χ_0 is a smooth point (see Section 5.1) of $X(\pi_1(N))$ – so that in particular there is only one irreducible component of $X(\pi_1(N))$ containing χ_0 – and that in a neighborhood of χ_0, the functions $I_{\gamma_1}, \ldots, I_{\gamma_i}$ form a system of local coordinates for X_0. This last statement could be used to simplify very slightly some of the arguments that I'll be giving later on. However, as this goes to press, it is not clear to me whether references for these stronger statements are available.

I'll conclude this subsection with a simplified statement of a special case of Theorem 4.5.1 which will come up in several applications. If $n = 1$, so that X_0 is a curve, and if we set $\gamma = \gamma_1$, saying that χ_0 is an isolated point of $X^* = X_0 \cap I_\gamma^{-1}(\{\pm 2\})$ boils down to saying that the polynomial function I_γ is nonconstant on X_0. So we may state:

COROLLARY 4.5.2. *Let N be an orientable one-cusped hyperbolic 3-manifold of finite volume. Then there is a 1-dimensional irreducible component X_0 of $X(\pi_1(N))$, containing the character of the lift of a representation associated to the hyperbolic structure of N, such that if γ is any nontrivial element of $\pi_1(N)$ carried by the boundary of the compact core of N, the function I_γ is nonconstant on X_0.*

5. Ideal points and trees

In this section I'll be talking about a construction that was first introduced by Culler and me in [20]. It gives a way of associating actions of a finitely generated group Γ on trees with "ideal points" (see Section 5.3 below) of a curve in the character space $X(\Gamma)$. This construction turns out to have various applications to 3-manifold theory, because $\mathrm{SL}(2, \mathbf{C})$-representations of π_1 of a connected 3-manifold are related to hyperbolic structures on the manifold (Section 1.6), whereas actions of π_1 on trees are related to essential surfaces (Sections 1.5 and 2). All the subsequent sections of the chapter will depend in some way on the construction I'll be describing here.

The construction depends on a little more background material from algebraic geometry than was used in Section 4.

5.1. *Some more algebraic geometry*

A point P of an algebraic set V is said to be *smooth of dimension d,* where $0 \leqslant d \leqslant N$, if P has a neighborhood U in $\mathbf{C}^N$ with the property that $V \cap U = Z \cap U$, where Z is the solution set of a system of $N - d$ polynomial equations

$$f_1(z_1, \ldots, z_N) = \cdots = f_{N-d}(z_1, \ldots, z_N) = 0,$$

the f_i being polynomials in the coordinates $z_1, \ldots, z_N$, such that the $N \times (N - d)$-matrix of partial derivatives

$$\left(\frac{\partial f_j}{\partial z_i}\right)_{1 \leqslant i \leqslant N,\ 1 \leqslant j \leqslant N-d}$$

at the point P is of rank $N - d$.

I pointed out in Section 4.2 that the irreducible components of an affine algebraic set are not in general disjoint from one another. However, it is not hard to show that a point lying in the intersection of two distinct components is not smooth in V, so a smooth point of V does lie in a unique irreducible component.

If V is an affine variety in $\mathbf{C}^N$, there is a unique natural number $d \leqslant N$, called the dimension of V, such that V has a dense subset consisting of smooth points of dimension d. If x is any smooth point of V, we can apply the complex implicit function theorem – which looks formally just like the real implicit function from advanced calculus, and can be deduced from it – to parametrize the points in some neighborhood of x by d complex coordinates. This gives the set of smooth points of d the structure of a complex submanifold of dimension d in $\mathbf{C}^N$.

An affine variety V of dimension 1 is called, naturally enough, an affine curve. In the 1-dimensional case, the existence of a local complex coordinate near a smooth point of V has especially nice consequences, for example for the study of the function field $\mathbf{C}(V)$. Using a local coordinate we can identify a neighborhood U of a smooth point $P \in V$ with a domain in $\mathbf{C}$ in such a way that $P = 0$. If we write a given element of $\mathbf{C}(V)$ in the form f/g, where $f, g \in \mathbf{C}[V]$, then the restrictions of $f, g \in \mathbf{C}[V]$ to U are holomorphic functions f_U and g_U. The quotient f_U/g_U is a meromorphic function, and takes a well-defined value in the Riemann sphere $\mathbf{C} \cup \infty$ at every point in its domain. More explicitly, if $f_U(z) = z^m F(z)$ and $g_U(z) = z^n G(z)$, where F and G are holomorphic and are nonzero at 0, then

$$\frac{f_U(z)}{g_U(z)} = z^{m-n} \frac{F(z)}{G(z)},$$

where F and G are holomorphic and nonzero. It's easy enough to show that the meromorphic function f_U/g_U is well-defined, i.e., does not depend on the way we wrote the given element of $\mathbf{C}(V)$ as a quotient. Its value at 0, which we may think of as the value of f/g at P, is of course $F(0)/G(0)$ if $m \geqslant n$ and ∞ if $m > n$.

In Section 4.2 I pointed out that an element of $\mathbf{C}(V)$ defines a natural complex-valued function on an open dense subset of V. What we are seeing here is a partial improvement of this: an element of $\mathbf{C}(V)$ defines a natural function on the set of all smooth points of V, although this function now takes values in $\mathbf{C} \cup \{\infty\}$. In any case, the value we have assigned to f/g at P is indeed the value in any reasonable sense; for example, it is the limit of $f(z)/g(z)$ as z approaches P through $U - \{P\}$.

The number $m - n$ that appeared in the discussion above is the *order* of f_U/g_U in the sense of Section 3.3. Now I pointed out in 3.3 that there is a valuation of the field of meromorphic functions on U which assigns to each function its order at 0. For the same reason, there is a valuation of the function field $\mathbf{C}(V)$ that assigns to each element f/g of $\mathbf{C}(V)$ the order at P of the corresponding function f_U/g_U on U.

For any point P of an arbitrary affine variety V, it is tempting to try to assign a value at P to any element of $\mathbf{C}(V)$, but this is not always possible. For example, if $V = \mathbf{C}^2$ (an affine space which we may think of as an algebraic subset of itself, defined by the empty set of equations), and if z and w denote the coordinates on $\mathbf{C}^2$, then the element z/w of $\mathbf{C}(V)$ defines a function on the complement of the line L defined by $w = 0$; furthermore, this function can be extended continuously to a map $\mathbf{C}^2 - \{(0,0)\} \to \mathbf{C} \cup \{\infty\}$ by giving it the value ∞ at every point of $L - \{(0,0)\}$. However, the function cannot be extended continuously to $(0,0)$, because it can take arbitrary limiting values through a sequence in $\mathbf{C}^2 - \{0,0\}$ which approaches $(0,0)$; for example, along the complement of $\{(0,0)\}$ in a line $z = \lambda w$, its value is identically equal to λ.

This problematic behavior can also occur at nonsmooth points of affine curves. For example, the two-variable polynomial $z^3 + w^3 + zw$ is easily seen to be irreducible, from which it follows by general principles that its zeros form an irreducible algebraic set of codimension 1, hence a curve, $V \subset \mathbf{C}^2$. Since the polynomial defining V is closely approximated by zw near $(0,0)$, one can show that the intersection of V with a suitable neighborhood of $(0,0)$ is made up of two "branches": these are complex analytic 1-manifolds that are tangent to the coordinate axes $z = 0$ and $w = 0$. Consider the element z/w of $\mathbf{C}(V)$, where z and w now denote the generators of $\mathbf{C}[V]$ obtained by restricting the coordinate functions to V. The "rational function" z/w defines a genuine function on $V - \{(0,0)\}$, but this function approaches 0 as the argument approaches $(0,0)$ through the branch tangent to $z = 0$, and approaches ∞ as the argument approaches $(0,0)$ through the branch tangent to $w = 0$.

5.2. *Projective varieties*

Affine varieties are often awkward to work with because they are noncompact. One can typically learn much more about an affine variety V by studying a *projective completion* of V.

Recall that for a positive integer N, the complex projective N-space $\mathbf{CP}^N$ is the quotient of $\mathbf{C}^{N+1} - \{(0,\dots,0)\}$ under the equivalence relation $\sim$ in which $(Z_0,\dots,Z_N) \sim (W_0,\dots,W_N)$ if and only if there is a complex number $\alpha \neq 0$ such that $W_i = \alpha Z_i$ for $i = 0,\dots,N$. I'll denote the equivalence class of $(Z_0,\dots,Z_N)$ by $[Z_0,\dots,Z_N]$. One says that $Z_0,\dots,Z_N$ are *homogeneous coordinates* for $[Z_0,\dots,Z_N]$.

If f is a homogeneous complex polynomial of degree $d \geqslant 0$ in $N+1$ indeterminates, then for any point $(Z_0,\dots,Z_N)$ of $\mathbf{C}^{N+1}$ and any $\alpha \in \mathbf{C}$ we have $f(\alpha Z_0,\dots,\alpha Z_N) =$

$\alpha^d f(Z_0, \ldots, Z_N)$. This means that although f does not have a well-defined value at a given point of $\mathbf{CP}^N$, it does have a well-defined set of zeros. A *projective algebraic set* in $\mathbf{CP}^N$ is the set of common zeros of a collection of homogeneous polynomials – of various degrees – in $N+1$ indeterminates. Some of the basic properties of affine algebraic sets have straightforward analogues for the projective case. Thus any projective variety can actually be defined as the zero set of a *finite* collection of homogeneous polynomials, and can be represented as a finite union of projective *varieties*, i.e., irreducible projective algebraic sets. (The definition of reducibility in the projective setting looks formally just like the affine definition.)

Let $H_0 \subset \mathbf{CP}^N$ denote the locus of zeros of the coordinate function Z_0, which we can think of as a first-degree homogeneous polynomial. The map $J_0 : \mathbf{C}^N \to \mathbf{CP}^N$ defined by $J_0(z_1, \ldots, z_N) = [1, z_1, \ldots, z_N]$ is a diffeomorphism of $\mathbf{C}^N$ onto the open dense subset $\mathbf{CP}^N - H_0$ of $\mathbf{CP}^N$, with inverse given by

$$[Z_0, Z_1, \ldots, Z_N] \mapsto \left(\frac{Z_1}{Z_0}, \ldots, \frac{Z_N}{Z_0}\right).$$

If V is any affine algebraic set in $\mathbf{C}^N$ then the closure $\overline{J_0(V)}$ is a projective algebraic set in $\mathbf{CP}^N$. If V is irreducible, so is $\overline{J_0(V)}$. (Sometimes we can find defining equations for $\overline{J_0(V)}$ by "homogenizing" equations for V: thus if $n = 2$ and V is defined by $z_1^2 + z_2^3 = 1$, we can define $\overline{J_0(V)}$ by $Z_1^2 + Z_2^3 = Z_0^3$. However, this will not always work. To give a trivial example, if we define $\emptyset \subset \mathbf{C}^2$ by the equations $z_1 = 1$, $z_1 = 2$, homogenizing gives the equations $Z_1 = Z_0$, $Z_1 = 2Z_0$. The solution set of the latter system consists of the point $[1, 1, 2]$, whereas $\overline{J_0(\emptyset)} = \emptyset$. What is always true is that if we homogenize an arbitrary system of equations defining V then $\overline{J_0(V)}$ is a union of irreducible components of the locus of zeros of the resulting homogeneous system.)

Because $\mathbf{CP}^N$ is obviously compact, projective varieties are always compact; this makes them more tractable objects than affine varieties for some purposes. If $V \subset \mathbf{C}^N$ is an affine variety, we can think of $\overline{J_0(V)}$ as a compactification of V, from which V can be "recovered" since $J_0(V) = \overline{J_0(V)} \cap J_0(\mathbf{C}^N)$. This is often a useful way of getting information about an affine variety.

On the other hand, we can also use affine varieties to study projective ones. If we're looking at a projective variety $W \subset \mathbf{CP}^N$, we can assume that W is not contained in any of the "hyperplanes" $H_i \subset \mathbf{CP}^N$ defined by $Z_i = 0$, for $i = 0, \ldots, N$, because otherwise we could think of W as a variety in a lower-dimensional projective space. Now if $J_0 : \mathbf{C}^N \to \mathbf{CP}^N - H_0$ is defined as above, and if V is an algebraic set not contained in H_0, then $V_0 = V \cap (\mathbf{CP}^N - H_0)$ is "identified" via J_0 with the set $J_0^{-1}(W)$; it's not hard to show that $J_0^{-1}(W)$ is an affine variety in $\mathbf{C}^N$. But for $i = 0, \ldots, n-1$, we can do exactly the same construction using H_i and a similarly defined map $J_i : \mathbf{C}^N \to \mathbf{CP}^N - H_i$ in place of H_0 and J_0, to get an "identification" of the set $V_i = V \cap (\mathbf{CP}^N - H_i)$ with an affine variety. We have $V = \bigcup_{i=0}^N V_0$, and we can think of the V_i as domains of an "atlas of affine coordinate charts"; in terms of these, a projective variety is something that looks "locally" like an affine variety, in much the way that a differentiable manifold looks locally like $\mathbf{R}^n$.

The "transition maps" are easily understood in the setting of a projective variety. To simplify the notation a bit, let's consider a point P lying in the intersection of the chart domains V_0 and V_1. Suppose that P, regarded as a point of the affine variety V_0, has affine coordinates $z_1, z_2, \dots, z_N$. Then $1, z_1, z_2, \dots, z_N$ are homogeneous coordinates for P as a point of $V \subset \mathbf{CP}^N$. It follows that if we regard P as a point of V_1, its affine coordinates are $1/z_1, z_2/z_1, \dots, z_N/z_1$. The same calculation shows that for any i and j, the "transition maps" relating affine coordinates in V_i to those in V_j are defined by rational functions in the coordinates.

In view of the description of the function field as a set of equivalence classes of functions, which I gave in Section 4.2, it now follows that there is a natural isomorphic identification of all the function fields $\mathbf{C}(V_0), \dots, \mathbf{C}(V_N)$ with one another. So it makes sense to talk about the *function field* $\mathbf{C}(W)$ of a projective variety W.

This is a first example of how one can turn "local" definitions involving affine varieties into "global" ones involving projective varieties, in close analogy with the theory of differentiable manifolds. There are of course many other examples. A point of a projective variety is *smooth* if it is identified with a smooth point of an affine variety under one of the affine charts; for a point lying in more than one chart domain, this is independent of the choice of chart, as a calculation with the transition maps shows. Likewise, a projective variety has a *dimension* which is equal to the dimension of each of its affine pieces.

The results about smooth points of affine curves that I discussed in Section 5.1 are readily translated into the projective context: if W is a projective curve, i.e., a projective variety of dimension 1, and if P is a smooth point of W, then every element of $\mathbf{C}(W)$ has a well-defined value at P, this value being an element of $\mathbf{C} \cup \{\infty\}$; and furthermore, P gives rise to a valuation of $\mathbf{C}(W)$ in a natural way.

5.3. *Canonical completions*

In Section 5.2 I talked about *the* completion of an affine algebraic set, because I was thinking of such sets concretely as subsets of particular affine spaces. As with so many other kinds of mathematical objects, it is often useful to have the flexibility that comes from thinking of affine algebraic sets as being "defined up to isomorphism". In this context it is no longer permissible to speak about *the* completion of an affine algebraic set, because isomorphic affine algebraic sets may have nonisomorphic projective completions. Actually, if you're paying close attention you'll have noticed that I haven't defined isomorphism of projective varieties, but there are examples where projective completions of isomorphic affine varieties are not even homeomorphic.

Let me point out an especially trivial example of this phenomenon. The union of two "parallel lines" in $\mathbf{C}^2$, say $z_1 = 0$ and $z_1 = 1$, is isomorphic as an affine algebraic set to the union of two "skew lines" in $\mathbf{C}^3$, say $z_1 = z_2 = 0$ and $z_2 = z_3 = 1$. (It's a good exercise to write down the polynomial maps between $\mathbf{C}^2$ and $\mathbf{C}^3$ that restrict to an isomorphism and its inverse.) On the other hand, the completion of the first set in $\mathbf{CP}^2$ is the union of the "projective lines" $Z_1 = 0$ and $Z_1 = Z_0$, which are topological 2-spheres meeting in the point $(0, 0, 1)$, whereas the completion of the second set is the union of the projective lines $Z_1 = Z_2 = 0$ and $Z_2 = Z_3 = Z_0$, which are again topological 2-spheres but are disjoint (since a point of $\mathbf{CP}^3$ cannot have all its homogeneous coordinates equal to 0). Note that in this example the completion of the second set is smooth, whereas the completion of the first set is not.

From this point on I will be using the phrase "completion of V", where V is an algebraic set in some $\mathbf{C}^N$, to mean any projective variety of the form $\overline{J_0(V')}$, where V' is an arbitrary affine algebraic set in some $\mathbf{C}^{N'}$ isomorphic to V and $J_0 : \mathbf{C}^{N'} \to \mathbf{CP}^{N'}$ is the standard embedding defined in Section 5.2. In situations where a particular completion $\widehat{V}$ of V has been fixed, I will regard V as being identified in the obvious way with a subset of $\widehat{V}$. In this situation I will often refer to the points of $\widehat{V} - V$ as *ideal points* of the completion $\widehat{V}$; in contrast the points of $V \subset \widehat{V}$ may be referred to as *ordinary points*.

The example I just described shows that a 1-dimensional algebraic set may have one completion in which the ideal points are smooth, and another which fails to have this property. It can be shown that every affine curve has a completion in which all the ideal points are smooth. This is in fact a fairly direct consequence of one of the basic results in the theory of algebraic curves, which allows one to "resolve the singularity" at a nonsmooth point of a projective algebraic curve, replacing it by a finite number of smooth points. In more precise terms, if x is a singular point of a projective curve C, there exist a projective curve $\widetilde{C}$ and a well-defined map $J : \widetilde{C} \to C$, which is rational in local affine coordinates near every point, such that $J^{-1}(y)$ is a single point for every $y \neq x$, and $J^{-1}(x)$ consists of a finite number of smooth points. In [48] and [31] you will find a total of three quite different proofs of this result, all very enlightening. Now if $\widehat{V}$ is any completion of an affine curve V, one can resolve all those singularities of $\widehat{V}$ which occur at ideal points, and it is not hard to show that the resulting projective curve is still a completion of V.

Although I will not be defining the notion of isomorphism of projective varieties in this chapter, because I won't really need it, I ought to mention that up to isomorphism there is only one completion of an affine curve in which all the ideal points are smooth. Once one has studied the definitions, the proof that such a completion is canonical in this sense is a simple application of the fact, which I talked about above, that if W is a projective curve, every element of $\mathbf{C}(W)$ has a well-defined value in $\mathbf{C} \cup \{\infty\}$ at every smooth point of W.

5.4. *Associating an action on a tree with an ideal point*

Let Γ be a finitely generated group, and let C be a curve contained in $X(\Gamma)$, i.e., an irreducible 1-dimensional subvariety of $X(\Gamma)$. By Section 5.3, there is a projective completion $\widehat{X}$ of X such that every ideal point of $\widehat{X}$ is smooth. I will show how every ideal point x of $\widehat{X}$ gives rise to a nontrivial action of Γ on a tree.

First of all, by the very way I defined $\mathbf{C}(\widehat{X})$ in Section 5.2, there is a natural isomorphism of $\mathbf{C}(\widehat{X})$ with $\mathbf{C}(X)$. From now on I'll write $F = \mathbf{C}(X) = \mathbf{C}(\widehat{X})$. Any ideal point x of $\widehat{X}$, because it is a smooth point, determines a valuation v_x of $F = \mathbf{C}(\widehat{X})$ by the construction described in Sections 5.1 and 5.2.

On the other hand, there is an irreducible subvariety R_C of $R(\Gamma)$ such that $t(R_C) = C$. (The point here is that since C is a subvariety of $X(\Gamma) = t(R(\Gamma))$, it is true for very general algebro-geometric reasons that C is the closure of $t(R_C)$ for some subvariety R_C of $R(\Gamma)$. The argument I alluded to in Section 4.4, based on the "Burnside Lemma", which shows that $t(R(\Gamma)) \subset \mathbf{C}^N$ is an algebraic set, also shows that $t(R_C)$ is an algebraic set, hence closed, hence equal to C. I'll have to refer you to [20] for details, but that's the philosophy.) As in Section 4.2, we can regard $K = \mathbf{C}(R_C)$ as an extension of the field $F = \mathbf{C}(X)$. We invoke the following extension theorem for valuations:

THEOREM 5.4.1. *Let K be a finitely generated extension of a field F and let $v: F^* \to \mathbf{Z}$ be a valuation of F. Then there exist an integer $d > 0$ and a valuation $w: K^* \to \mathbf{Z}$ such that $w|F^* = dv$.*

This result is pretty well-known and elementary. The best reference I can give you is to my joint paper [2] with Roger Alperin, where we state it as Lemma 1.1 and give a proof that's self-contained except for a reference to Bourbaki's Commutative Algebra. If you read the proof and look up the reference, you'll know about as much valuation theory as you need for this subject. The result is an extension theorem in the sense that the function $\frac{1}{d}w$ is an extension of v to K^*, and is a valuation in a very slightly more general sense than the one I have defined here: it takes its values in the infinite cyclic group $\frac{1}{d}\mathbf{Z}$ rather than in $\mathbf{Z}$.

We can apply this theorem in our situation because K, being the function field of a (finite-dimensional) variety R_C, is finitely generated as an extension of $\mathbf{C}$; in fact, if C lives in an affine space $\mathbf{C}^N$ then the restrictions of the coordinate functions to R_C generate K over $\mathbf{C}$. So in particular K is finitely generated as an extension of F. The theorem gives a valuation w of F such that $w|F^* = dv_x$ for some $d > 0$. With the valuation w, as in Section 3, we can associate a tree $T = T_w$ on which $\mathrm{SL}(2, K)$ acts in a natural way. On the other hand, by Section 4 we have a tautological representation $\mathcal{P}: \Gamma \to \mathrm{SL}(2, K)$, and we can pull back the action of $\mathrm{SL}(2, K)$ on T via $\mathcal{P}$ to get an action of Γ on T. According to Section 3.7, Γ acts without inversions on T.

For any $\gamma \in \Gamma$, the function I_γ is an element of $\mathbf{C}(X)$. Since the ideal point x is smooth, it follows from what I said in Section 5.2 that I_γ has a well-defined value $I_\gamma(x) \in \mathbf{C} \cup \infty$ at x. The most important property of the action of Γ on the tree T is:

PROPERTY 5.4.2. *For any element $\gamma \in \Gamma$ the following statements are equivalent:*

(i) $I_\gamma(x) \in \mathbf{C}$, *i.e., I_γ does not have a pole at x;*

and

(ii) *some vertex of T is fixed by γ.*

To prove this, first recall that by (4.4.1) the element I_γ of $F \subset K$ is the trace of $\mathcal{P}(\gamma) \in \mathrm{SL}(2, K)$. We have

$$I_\gamma(x) \in \mathbf{C} \Leftrightarrow v(I_\gamma) \geqslant 0 \Leftrightarrow w(\operatorname{trace} \mathcal{P}(\gamma)) \geqslant 0 \Leftrightarrow \operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}, \tag{5.4.3}$$

where $\mathcal{O} \subset K$ is the valuation ring defined by the valuation w. Now if (ii) holds, i.e., if $\mathcal{P}(\gamma) \in \mathrm{SL}(2, K)$ fixes a vertex of T, then by Section 3.7 the element $\mathcal{P}(\gamma)$ lies in a conjugate, within $\mathrm{GL}(2, K)$, of $\mathrm{SL}(2, \mathcal{O})$. In particular, $\operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}$, so that (i) holds by virtue of (5.4.3). To prove the converse we need the *rational canonical form* of a matrix in $\mathrm{SL}(2, K)$, which in this case is a pretty trivial matter. Suppose that (i) holds, so that $\operatorname{trace} \mathcal{P}(\gamma) \in \mathcal{O}$. If we are in the degenerate case where $\mathcal{P}(\gamma) = \pm I$ then γ acts trivially on the whole tree T. In the nondegenerate case we can choose a vector $e \in K^2$ such that e and its image f under the linear transformation $A = \mathcal{P}(\gamma)$ of K^2 are linearly independent. In the basis $\{e, f\}$, the linear matrix of the linear transformation A has the form

$$B = \begin{pmatrix} 0 & 1 \\ c & d \end{pmatrix},$$

so that A is conjugate to B in $\mathrm{GL}(2, K)$. In particular we have $c = -\det A = -1$ and $d = \operatorname{trace} A \in \mathcal{O}$, so that $B \in \mathrm{SL}(2, \mathcal{O})$. Thus $A = \mathcal{P}(\gamma)$ lies in a conjugate, within $\mathrm{GL}(2, K)$, of $\mathrm{SL}(2, \mathcal{O})$, and by Section 3.7, γ fixes a vertex of T.

An important consequence of Property 5.4.2 is the following property of the action:

PROPERTY 5.4.4. *The action of Γ on T is nontrivial.*

(Recall from Section 1.5 that this means that no vertex of T is fixed by the entire group Γ.) To prove this, note that if the action were trivial, then by Property 5.4.2 we would have $I_\gamma(x) \in \mathbf{C}$ for every element γ of Γ. In terms of the concrete description of the character variety that I gave in Section 4.4, each of the coordinate functions of $C \subset X(\Gamma)$ is of the form I_γ for some $\gamma \in \Gamma$. But some coordinate function must take the value ∞ at x since x is an ideal point. This contradiction completes the proof of Property 5.4.4.

5.5. *More about the action*

In some of the applications that I discuss in this chapter I will need an apparently more general version of Property 5.4.2: a finitely generated subgroup Γ_1 of Γ fixes a vertex of T if and only if I_γ takes a finite value at x for every $\gamma \in \Gamma$. My favorite way to prove this is to notice that it follows immediately from Property 5.4.2 itself and the following result:

PROPOSITION 5.5.1. *Suppose that a finitely generated group Γ acts without inversions on a tree T, in such a way that each element $\gamma \in \Gamma$ fixes some vertex v_γ of T. Then there is a single vertex v of T which is fixed by the entire group Γ: thus $\gamma \cdot v = v$ for every $\gamma \in \Gamma$.*

You may find a proof of this in [58], but it's much more fun to do it as an exercise.

Property 5.4.2 also admits a generalization in a different direction. This version involves the notion of the *length function* associated to an action of a group Γ on a tree T. The length function l associated to an action $\cdot$ (without inversions) is defined by $l(\gamma) = \min_s d(s, \gamma \cdot s)$, where s ranges over the vertices of T. I mentioned in Section 3.8 that any $\gamma \in \Gamma$ either has a fixed point in T – in which case $l(\gamma) = 0$ – or has a unique invariant line (an "axis") on which it acts by a translation; in this case it is not hard to show that $l(\gamma)$ is the (integer) distance through which γ translates vertices on its axis.

The second generalization of 5.4.2, involving the length functions l associated to the action of Γ on T, is that for any $\gamma \in \Gamma$, the length $l(\gamma)$ is equal to twice the order of the pole of I_γ at the ideal point x. (Here "the order of the pole" is taken to mean 0 if I_γ does not have a pole at x, so that 5.4.2 indeed appears as a special case.) This is also an excellent exercise. The only reference I know for it is [45], where it is proved in a much more general form.

One consequence of this generalization of 5.4.2 is that the length function defined by the action of Γ on $T = T_w$ is canonically associated to the ideal point x, i.e., does not depend on the choice of the extension w.

The length function associated to an action on a tree plays the same role as the character associated to a representation in SL(2, **C**), and there is an analogue of Proposition 1.1.1. To understand the statement, first note that according to Proposition 5.5.1, an action of a finitely generated group Γ on a tree is trivial if and only if the associated length function is 0. It's not hard to show that if Γ acts nontrivially on T then there is a unique minimal Γ-invariant subtree of T. (It can be described as the union of the axes of all the elements of Γ that do not have fixed vertices.) It's easy to see that if two actions of Γ on trees have minimal invariant subtrees that are equivariantly simplicially isomorphic, then they give rise to the same length function. The analogue of Proposition 1.1.1 gives a converse that is valid except for certain "degenerate" actions that are analogous to reducible representations.

An action of Γ on a tree is termed *abelian* if its length function has the form $l(\gamma) = |h(\gamma)|$ where h is a homomorphism from Γ to **Z**; any function of this form does arise from an action of Γ on the tree **R**, triangulated so that its vertex set is **Z**. The following result is a special case of results proved in [19] and in [1]:

PROPOSITION 5.5.2. *Let Γ be a finitely generated group. If two nonabelian actions of Γ on trees T and T' define the same length function, then the minimal Γ-invariant subtrees of T and T' are Γ-equivariantly simplicially isomorphic.*

In particular, for the case of the action of Γ on the tree T associated to an ideal point by the construction I described above, the restriction of the action to the minimal Γ-invariant subtree of T is something canonically defined by the given ideal point – except in the degenerate case where the action is abelian. I won't be using this fact anywhere in the rest of this chapter, but as Golde and Tevye said, it's nice to know.

I'll conclude this section by pointing out one more property of the action of Γ on the tree T associated to an ideal point x, which is just the translation of (3.8.2) into this context. (Check it.) This property and its variants come up a lot in applications to 3-manifolds.

PROPERTY 5.5.3. *If e is any edge of T_w, and γ is any element of $[\Gamma_e, \Gamma_e]$, the commutator subgroup of the stabilizer of e in Γ, then $I_\gamma(x) = 2$.*

5.6. *Separating surfaces in knot exteriors*

In his classic treatise [50], Lee Neuwirth asked a number of questions about the structure of knot groups, i.e., fundamental groups of complements of nontrivial knots in S^3. (A knot is said to be *trivial* if it bounds a disk in S^3.) One of his questions was whether every knot group can be expressed as a nontrivial free product with amalgamation (see Section 2.6) in which the amalgamated subgroup is free. He proposed the idea of answering this question affirmatively by showing that if K is a nontrivial tame knot in S^3 then the exterior of K contains a separating essential surface. I will refer to this statement as the *weak Neuwirth Conjecture*, because in [49] Neuwirth formulated stronger topological and group-theoretic versions of his conjecture, some of which are still unproved. Of course you should compare the weak Neuwirth Conjecture with the elementary fact, which I talked about in Section 2.5, that the exterior of K always contains a nonseparating essential surface.

The weak Neuwirth Conjecture was proved in [21], in a much more general context than that of knot exteriors in S^3. It is included in the following result, which is proved in [21]:

THEOREM 5.6.1 (Culler–Shalen). *Let M be a compact, orientable, irreducible 3-manifold whose boundary is a torus. Suppose that $H_1(\partial M; \mathbf{Q}) \to H_1(M; \mathbf{Q})$ is surjective, but that M is not a solid torus. Then M contains a separating essential surface.*

(By the way, when M is the exterior of a tame knot in S^3, the surjectivity of $H_1(\partial M; \mathbf{Q}) \to H_1(M; \mathbf{Q})$ follows easily from the Mayer–Vietoris theorem. The irreducibility of M in this case follows from a classical result due to Alexander, Graeub and Moise, the so-called 3-dimensional PL Schönflies theorem (see, for example, [42, Section 17]).)

I will discuss the stronger versions of Neuwirth's conjectures in Section 10. The essential point to be made here is that while Theorem 5.6.1 applies to irreducible tame knot exteriors in arbitrary rational homology 3-spheres, the stronger versions of the conjecture seem to depend on the hypothesis that the knot is in S^3, or at any rate in a 3-manifold of some more special sort.

Actually Theorem 5.6.1 is an essentially immediate consequence of a result, stated below as Theorem 5.6.2, which applies to irreducible tame knot exteriors in arbitrary closed, orientable, connected 3-manifolds; the statement of this theorem was not spelled out in [21]. To state it we need the notion of a *boundary slope*, which is discussed in more detail in Boyer's chapter in this Handbook [7]. Briefly, if M is a compact, orientable, irreducible 3-manifold whose boundary is a torus, the boundary components of an essential surface F are all isotopic since they are disjoint, homotopically nontrivial simple closed curves on a torus. Their common isotopy class is called the *boundary slope* of F; for reasons that Boyer explains, the term *slope* is used to mean any isotopy class of nontrivial simple closed curves in ∂M. A slope is called a boundary slope of M if it is the boundary slope of some essential surface in M. A fundamental result of Hatcher's [33] implies that the set of boundary slopes of M is always finite.

THEOREM 5.6.2. *Let M be a compact, orientable, irreducible 3-manifold whose boundary is a torus. Then either*
(i) *M is a solid torus, or*
(ii) *M contains an essential separating annulus, or*
(iii) *M contains an essential nonseparating torus, or*
(iv) *M has at least two boundary slopes.*

To see that this implies Theorem 5.6.1, notice that alternative (i) is ruled out by the hypothesis of Theorem 5.6.1, while alternative (ii) implies the conclusion of 5.6.1. The hypothesis that $H_1(\partial M; \mathbf{Q}) \to H_1(M; \mathbf{Q})$ is surjective rules out alternative (iii). This hypothesis also implies, by a simple homological argument, that there is at most one slope that can occur as the boundary slope of a nonseparating essential surface. So if alternative (iv) holds then M contains an essential separating surface.

In this subsection I will show how to apply the techniques that I've described to prove Theorem 5.6.2 in the special case where $N = \operatorname{int} M$ has a (finite-volume) hyperbolic structure. In this case one gets a stronger result, namely that alternative (iv) of Theorem 5.6.2 holds. I'll establish alternative (iv) in the following paraphrased form: if s is any slope, there is an essential surface which has nonempty boundary and has a boundary slope different from s.

Let s be represented by a simple closed curve c in ∂M and let γ be an element representing the conjugacy class determined by some orientation of c. We consider the curve $X_0 \subset X(\pi_1(M))$ given by Corollary 4.5.2. Since I_γ is nonconstant on X_0 it must have a pole on $\widehat{X}_0$, which must occur at some ideal point x of X_0.

With the ideal point x we can associate a tree T_x and an action of $\pi_1(M)$ on T_x by the construction of Section 5.4. Furthermore, according to Section 2 there exists an essential surface $F \subset M$ which is dual to the action of $\pi_1(M)$ on T_x. To complete the proof it suffices to show that any such F has a nonempty boundary and that its boundary slope is different from s.

Assume that either F is closed, or that $\partial F \neq \emptyset$ and that the boundary slope of F is s. In either case, the simple closed curve c is isotopic to a simple closed curve in the complement of F. Thus the element $\gamma \in \pi_1(M)$ lies in a conjugate of $\operatorname{im}(\pi_1(C_i)) \to \pi_1(M)$ for some component C_i of $M - F$. According to 2.3.1(i), this implies that γ fixes some vertex of T_x. But by Property 5.4.2 this means that I_γ does not have a pole at x. Of course this contradicts our choice of x, and so the proof of the theorem is complete – in the special case where $\operatorname{int} M$ is hyperbolic.

6. The proof of the weak Neuwirth Conjecture

I'm going to present the main ideas in the proof of Theorem 5.6.2 in the general case, where we don't assume that $\operatorname{int} M$ has a hyperbolic structure. I will skip over a few technical algebraic details, which you can find in [21].

The starting point for the argument is Thurston's geometrization theorem [52,53], which, together with the characteristic submanifold theorem [38,37] guarantees that by splitting M along some disjoint system of essential tori $\{T_1, \dots, T_k\}$ we can get a manifold M' such that for each component M'_i of M', either M'_i is a Sefiert fibered space or $\operatorname{int} M'_i$ is hyperbolic. We can assume that the component M_0 of M' which contains the torus $B_0 = \partial M$ is not homeomorphic to $S^1 \times S^1 \times [0, 1]$, as otherwise we could replace $\{T_1, \dots, T_k\}$ by a system of fewer tori with the same properties.

If M_0 is a Seifert fibered space (not homeomorphic to $S^1 \times S^1 \times [0, 1]$), it's a routine matter to check that either M_0 is homeomorphic to $D^2 \times [0, 1]$ – in which case $M = M_0$ and alternative (i) of Theorem 5.6.2 holds – or M_0 contains an essential separating annulus A with $\partial A \subset B_0$. We can think of A as an essential annulus in M. This annulus will separate M (implying alternative (ii) of Theorem 5.6.2) unless there are two components T_1 and T_2 of ∂M_0 which lie in different components of $M_0 - A$ but lie in the same component of $M - M_0$; but if this happens then T_1 and T_2 are both nonseparating tori in M, so that alternative (iii) of Theorem 5.6.2 will hold. So the conclusion of the theorem holds whenever M_0 is a Seifert fibered space.

In the crucial case where int M_0 is hyperbolic, we generalize the argument of Section 5.6 to show that alternative (iv) of Theorem 5.6.2 holds. As I pointed out in Section 5.6, it suffices to show that if s is any slope, there is an essential surface in M which has nonempty boundary and has a boundary slope different from s. It turns out we can show more than this, namely that there is an essential surface $F \subset M_0$ with $\partial F \subset B_0$, and such that the common isotopy class of the components of ∂F is distinct from s.

In Section 5.6, in the case where int M was hyperbolic and had a single cusp, we used the 1-dimensional component of $\pi_1(M)$ given by Corollary 4.5.2. In the general case, if int M_0 has n cusps then ∂M_0 consists of n tori. We have labeled one of these B_0; let $B_1, \dots, B_{n-1}$ denote the others. Theorem 4.5.1 gives an n-dimensional irreducible component X_0 of $X(\Gamma)$. I'll define a curve $Y_0 \subset X_0$ which plays the role that X_0 played in the one-cusp case. Specifically, I claim that X_0 contains a curve Y_0 such that

(i) for each $i = 1, \dots, n-1$ and each element $\alpha \in \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$, the function $I_\alpha | Y_0$ is identically equal to either 2 or -2, and

(ii) for each nontrivial element $\alpha \in \mathrm{im}(\pi_1(B_0) \to \pi_1(M))$, the function $I_\alpha | Y_0$ is non-constant.

To construct Y_0 we first recall that X_0 is by definition an irreducible component of $X(\pi_1(M))$ containing the character χ_0 of the lift to SL(2, **C**) of a discrete faithful representation of $\pi_1(M)$ in PSL(2, **C**). Hence if we fix elements $\gamma_i \in \mathrm{im}(\pi_1(B_i) \to \pi_1(M))$ for $i = 1, \dots, n-1$, we have $I_{\gamma_i}(\chi_0) = \chi_0(\gamma_i) = \pm 2$ for $1 \leqslant i \leqslant n-1$. In particular, χ_0 lies in the algebraic subset Z of $X(\pi_1(N))$ obtained by adding the equations $I_{\gamma_i}^2 = 4$ for $i = 1, \dots, n-1$. Now by a general property of complex affine varieties that I already quoted in Section 4.5 (see [59, p. 60, Theorem 7]), since X_0 has dimension n, and Z is defined by adding $n-1$ extra equations, each component of Z must have dimension at least 1. In particular there must be a curve Y_0 with $\chi_0 \in Y_0 \subset X_0$. To prove property (ii) for the curve Y_0, we use the fact that $I_\alpha(\chi_0) = \pm 2$ to conclude that if I_α were constant on Y_0, all the functions $I_\alpha^2, I_{\gamma_1}^2, \dots, I_{\gamma_{n-1}}^2$ would be identically equal to 4 on Y_0. Thus Y_0 would be contained in the algebraic subset X^* of $X(\pi_1(N))$ obtained by adding the equations $I_\alpha^2 = I_{\gamma_1}^2 = \dots = I_{\gamma_{n-1}}^2 = 4$. But this contradicts the last assertion of Theorem 4.5.1, according to which χ_0 is an isolated point of X^*.

To show that Y satisfies (i), one begins with the fact, which I already mentioned in Section 5.4 (the paragraph before the statement of the Extension Theorem for Valuations) that any (irreducible) curve in $X(\pi_1(M))$ is in fact the image under $t : R(\pi_1(M)) \to X(\pi_1(M))$ of a subvariety of $R(\pi_1(M))$. Having fixed a subvariety R_0 of $R(\pi_1(M))$ with $t(R_0) = Y_0$, one considers an index i with $1 \leqslant i < n$ and an element α of $\mathrm{im}(\pi_1(B_i) \to \pi_1(M))$. For any point $\rho \in R_0$ we have trace $\rho(\gamma_i) = I_{\gamma_i}(t(\rho)) = \pm 2$, since $t(\rho) \in Y_0 \subset Z$. Thus $\rho(\gamma_i)$ is either $\pm I$ or a conjugate of $\pm\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. But $\rho(\alpha)$ commutes with $\rho(\gamma_i)$ since $\pi_1(B_i)$ is abelian; as the only matrices that commute with $\pm\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ are those of the form $\pm\begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$ for $\lambda \in \mathbf{C}$, we must have either $\rho(\gamma_i) = \pm I$ or trace $\rho(\alpha) = \pm 2$.

In view of the irreducibility of R_0, there are two possibilities: either (a) $\rho(\alpha_i) = \pm I$ for *every* $\rho \in R_0$, or (b) trace $\rho(\alpha) = \pm 2$ for every $\rho \in R_0$. Now (b) is exactly what we need, because it translates into the statement that the function I_α takes the value 2 or -2 at every point of Y_0; this implies the conclusion of (i) since Y_0 is irreducible. So we need only rule out (a). Well, if (a) holds then every point of X_0, in particular $\chi_0 = t(\rho_0)$, is the character of a nonfaithful representation. But since ρ_0 is irreducible, any representation with character χ_0 is equivalent to ρ_0, and since ρ_0 is faithful we have a contradiction.

Having constructed the curve Y_0 satisfying (i) and (ii), we proceed to construct the required surface F. We follow the same basic procedure as in Section 5.6, using Y_0 in place of X_0. Remember that we are given a slope s in B_0, and that we require that the essential surface $F \subset M_0$ have its boundary contained in B_0, and that the common isotopy class of the components of ∂F be distinct from s. We represent s by a simple closed curve c in ∂M, and we let γ be an element representing the conjugacy class determined by some orientation of c. By property (ii) of the curve Y_0, the function $I_\alpha | Y_0$ is nonconstant, and therefore has a pole at some ideal point x of $\widetilde{Y}_0$. The ideal point x determines an action of $\pi_1(M)$ on a tree T. If F is any essential surface in M dual to this action, the same argument that we used in Section 5.6 shows that F must have boundary components contained in B_0, and that the common isotopy class of these boundary components cannot be s. The new twist is that we must pick the surface F in such a way that it has no boundary components in any component $B_i \neq B_0$ of ∂M. This is made possible by property (i) of Y_0; I'll indicate how this works.

First we translate property (i) into a property of the action of $\pi_1(M)$ on T. For $i = 1, \ldots, n-1$, set $\Gamma_i = \operatorname{im}(\pi_1(B_i) \to \pi_1(M))$. Of course, this subgroup is defined only up to conjugacy, but we fix a concrete representative Γ_i in the conjugacy class of subgroups. Property (ii) says that I_γ takes a finite value at the ideal point x for every $\gamma \in \Gamma_i$. Thus by Proposition 5.5.1, there is a vertex v_i of T which is fixed by Γ_i.

The property of the action that I have stated here is exactly what is needed to guarantee that we can choose a surface F dual to the action in such a way that F is disjoint from $B_1, \ldots, B_{n-1}$. (Remember that the surface dual to a given action is in general far from being canonical.) The transition from the property of the action to the property of the (suitably chosen) dual surface is contained in the following result, which is a corollary to Proposition 2.5.3.

COROLLARY 6.0.1. *Let M be a compact, orientable, irreducible 3-manifold, and let $B_1, \ldots, B_k$ be disjoint subpolyhedra of ∂M. Suppose that we are given an action of $\pi_1(M)$ on a tree T, without inversions, and suppose that for each $i \leqslant k$, the subgroup $\Gamma_i = \operatorname{im}(\pi_1(B_i) \to \pi_1(M))$ of $\pi_1(M)$ fixes a vertex v_i of T. (Of course the subgroups Γ_i are defined only up to conjugacy, but if a given subgroup fixes a vertex then any conjugate subgroup fixes a (possibly different) vertex; so the condition makes sense.) Then there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap B_i = \emptyset$ for $i = 1, \ldots, k$.*

This is just the special case of Proposition 2.5.3 in which the map $\tilde{g}_i$, for $i = 1, \ldots, k$, is the constant map that sends $\widetilde{B}_i$ to v_i; it has the required equivariance property because Γ_i fixes v_i. In the notation of 2.5.3 we have $C_i = \emptyset$ for $i = 1, \ldots, k$, so that the conclusion of 2.5.3 implies that $F \cap B_i = \emptyset$.

In the case of the action associated to the ideal point x that we are considering, I have shown that the hypothesis of Corollary 6.0.1 holds if we set $k = n - 1$ and define

$B_1, \ldots, B_{n-1}$ as above. So the corollary gives an essential dual surface F to the action which has the property that $F \cap B_i = \emptyset$ for $i = 1, \ldots, n-1$. This completes the proof of Theorem 5.6.2.

7. The Smith Conjecture

It is a long-standing conjecture that a tame periodic homeomorphism h of S^3 is topologically linear, i.e., conjugate to the restriction to S^3 of an orthogonal transformation of $\mathbf{R}^4$. (To say that h is *tame* means that for every fixed point P of h, there exist an h-invariant neighborhood V of h and a homeomorphism j of V onto the unit ball $B^3 \subset \mathbf{R}^3$, such that jhj^{-1} is the restriction of a linear automorphism of $\mathbf{R}^3$. A periodic *diffeomorphism* is automatically tame.) The Smith Conjecture, which was first proved in 1978 (see [44]), is equivalent to the special case of this conjecture in which h is assumed to have period $n > 1$ (not much of a restriction), to preserve orientation, and to have nonempty fixed point set. Under these assumptions, it follows from classical theorems due to P.A. Smith that the fixed point set of h is homeomorphic to a 1-sphere, so that it may be regarded as a knot in S^3; and the hypothesis that h is tame implies that the fixed point set is a tame knot.

It was long unknown whether this fixed point set could be a nontrivial knot. In fact, classical arguments may be used to show that the homeomorphism is topologically linear if and only if the fixed point set is unknotted. This connection with classical knot theory was one source of fascination with the Smith Conjecture. However, the proof that was eventually given did not use this particular reduction to knot theory, but another one which in a sense is more direct.

If h is a tame, periodic homeomorphism whose fixed point set is a knot, it's elementary to show that the orbit space $\Sigma = S^3/\langle h \rangle$ is a connected, closed, orientable 3-manifold; and, furthermore, that the orbit map $p: S^3 \to \Sigma$ is a branched covering map, whose branch set is a tame knot $K \subset \Sigma$. It's pretty clear that the branched covering S^3 of Σ is regular and that its covering group is the cyclic group $\langle h \rangle$. Furthermore, it's an entirely elementary exercise to prove that h is topologically linear if and only if the knot $K \subset \Sigma$ is trivial, in the sense that it bounds a disk in Σ. So the Smith Conjecture is in fact an immediate consequence of the following result.

THEOREM 7.0.1. *Let K be a tame knot in a connected, closed, orientable 3-manifold Σ. Let n be an integer > 1, and let $\widetilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Assume that $\widetilde{\Sigma}$ is simply connected. Then the knot K is trivial.*

Of course this result is really stronger than the Smith Conjecture, because it is assumed only that $\widetilde{\Sigma}$ is simply connected, not that it is diffeomorphic to S^3. For this reason the result was known for a time as the "Generalized Smith Conjecture", until a still more general version was announced by Thurston ....

In this section I will sketch a proof of Theorem 7.0.1 which was first given in [20]. As I will point out below, it gives a rather stronger form of Theorem 7.0.1 than the form that was first proved in [44]. In any case, it is a good illustration of the use of the character variety in this subject.

7.1. *Preliminary observations*

A preliminary step in the proof of Theorem 7.0.1 is to reduce it to the special case where the exterior M of K is irreducible. This may be done by using the Kneser finiteness theorem – of which you will find an account in Bonahon's chapter in this volume, or in [34] (Theorem 3.15) – to decompose M as the connected sum of an irreducible manifold, itself the exterior of some tame knot K_0 in a closed orientable 3-manifold Σ_0, and a closed manifold. The n-fold cyclic branched cover $\widetilde{\Sigma}_0$ of Σ branched over K_0 is then a connected summand of $\widetilde{\Sigma}$, so if $\widetilde{\Sigma}$ is simply connected, so is $\widetilde{\Sigma}_0$. It's also a routine matter to check that K is trivial if and only if K_0 is.

There is a basically similar preliminary reduction to the case where K is a *prime* knot. A tame knot is said to be *prime* if its exterior contains no essential annulus whose boundary components are meridians. (I'm using the term "essential" as I did in Section 1, and since at this point I'm looking at a knot whose exterior is irreducible, the context is consistent with that of Section 1.) The reason for the term "prime" is that if the exterior of K does contain an essential annulus with meridian boundaries, we can decompose K as a *connected sum* of two tame knots K_1 and K_2 in closed orientable 3-manifolds Σ_1, Σ_2. (The connected sum is defined by removing from each Σ_i a ball B_i that meets K_i in an unknotted arc α_i, then gluing together $\Sigma_1 - \operatorname{int} B_1$ and $\Sigma_2 - \operatorname{int} B_2$ by some homeomorphism of their boundaries that matches $\partial\alpha_1$ with $\partial\alpha_2$; this gives a knot $K_1 \# K_2$ in the manifold $\Sigma_1 \# \Sigma_2$. For $\Sigma_1 = \Sigma_2 = S^3$, this formalizes the idea of tying two knots in succession in the same piece of string.) There is a prime decomposition of a tame knot analogous to the prime decomposition for an oriented 3-manifold described in Bonahon's chapter, and using this it is not hard to reduce the proof of Theorem 7.0.1 to the case where K, in addition to having an irreducible exterior, is prime.

Having made these reductions, we consider a prime tame knot, with irreducible exterior M, in a connected, closed, orientable 3-manifold Σ, and an integer $n > 1$. We let $\widetilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Working contrapositively, we assume that K is a nontrivial knot, and we wish to show that $\pi_1(\widetilde{\Sigma})$ is a nontrivial group. We let μ denote the meridian μ of K, which we think of as an element of $\pi_1(M)$, represented by a simple closed curve in ∂M. When I need to refer to this simple closed curve itself, regarded as a subset of ∂M, I will denote it $|\mu|$. It is quite elementary to see that there is an exact sequence

$$1 \to \pi_1(\widetilde{\Sigma}) \to \left|\pi_1(M)\colon\ \mu^n = 1\right| \to \mathbf{Z}/n\mathbf{Z} \to 0.$$

Here by $|\pi_1(M)\colon\ \mu^n = 1|$ I of course just mean the group obtained from $\pi_1(M)$ by adding the relation $\mu^n = 1$. (Nowadays this group is often referred to as an orbifold group.)

7.2. *The character variety appears*

Thus one way of showing that $\pi_1(\widetilde{\Sigma})$ is nontrivial is to show that $|\pi_1(M)\colon\ \mu^n = 1|$ is not a cyclic group. Thus it certainly suffices to find a representation $\bar{\rho} : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ with noncyclic image such that $\bar{\rho}(\mu)$ has order n. To carry this a step further, it suffices to find a representation $\rho : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ such that

(i) the image of $\rho(\pi_1(M))$ under the natural homomorphism $\mathrm{SL}(2,\mathbf{C}) \to \mathrm{PSL}(2,\mathbf{C})$ is noncyclic, and

(ii) $\rho(\mu)$ has order $2n$.

(I am using here the fact that $-I$ is the only element of order 2 is $\mathrm{SL}(2,\mathbf{C})$, so that any element of order $2n$ in $\mathrm{SL}(2,\mathbf{C})$ maps to an element of order n in $\mathrm{PSL}(2,\mathbf{C})$.)

Notice that I haven't claimed that there is always a representation ρ satisfying (i) and (ii). What I'll show, rather, is that in the crucial case where $\operatorname{int} M \cong \Sigma - K$ is hyperbolic (i.e., has a hyperbolic structure of finite volume), the attempt to find a representation ρ satisfying (i) and (ii) either succeeds or – in the case where it fails – leads to an alternative way of showing that $\pi_1(\widetilde{\Sigma})$ is nontrivial. It will also turn out that when $\operatorname{int} M$ is not hyperbolic, this alternative method nearly always works – and the exceptional cases are easily handled.

Assume for the moment, then, that $\operatorname{int} M$ is hyperbolic. Corollary 4.5.2 gives a curve $X_0 \subset X(\pi_1(M))$. We are attempting to find a point $\chi \in X_0$ which is the character of a representation ρ satisfying (i) and (ii). For condition (i) the following lemma is of obvious relevance.

LEMMA 7.2.1. *Let N be an orientable hyperbolic 3-manifold of finite volume with a single cusp, and let X_0 be a curve in $X(\pi_1(N))$ given by Corollary 4.5.2. Then for every point $\chi \in X_0$ there is a representation $\rho \in t^{-1}(\chi)$ such that the image of $\rho(\pi_1(N))$ under the natural homomorphism $\mathrm{SL}(2,\mathbf{C}) \to \mathrm{PSL}(2,\mathbf{C})$ is noncyclic.*

PROOF (*sketch*). We have $X_0 = t(R_0)$, where R_0 is an irreducible component of $R(\pi_1(N))$ containing ρ_0, a lift of the discrete, faithful representation of $\pi_1(N)$ to $\mathrm{SL}(2,\mathbf{C})$. It's easy to see that ρ_0 is irreducible. It's also easy to see that the reducible representations of $\pi_1(N)$ form a closed algebraic subset of $R(\pi_1(N))$. (In fact, it's shown in [20] that a representation ρ is reducible if and only if $\operatorname{trace}\rho(\gamma) = 2$ for every element γ of the commutator subgroup of $\pi_1(N)$.) So there's an open, dense subset of R_0 consisting of irreducible representations.

If $\rho \in R_0$ is irreducible then $\rho(\pi_1(N))$ is nonabelian. This easily implies that the centralizer of $\rho(\pi_1(N))$ is $\{\pm I\}$. Hence under the action of $\mathrm{SL}(2,\mathbf{C})$ by conjugation on R_0, the orbit $\rho^{\mathrm{SL}(2,\mathbf{C})}$ of ρ – which for the irreducible representation coincides with the fiber $t^{-1}(t(\rho))$ of ρ – is homeomorphic to $\mathrm{PSL}(2,\mathbf{C})$ and thus has dimension 3. Since X_0 has dimension 1 and the generic fiber of $t|R_0 : R_0 \to X_0$ has dimension 3, it follows that the dimension of R_0 is 4. By a general property of algebraic maps between complex varieties, it now follows that *every* fiber of $t|R_0$ has dimension at least $\dim R_0 - \dim X_0 = 3$.

Suppose now that there is a point $\chi \in X_0$ such that for every representation $\rho \in t^{-1}(\chi)$, the image of $\rho(\pi_1(N))$ in $\mathrm{PSL}(2,\mathbf{C})$ is cyclic. If for some given $\rho \in t^{-1}(\chi)$ the group $\rho(\pi_1(N))$ is not itself cyclic, then $\rho(\pi_1(N))$ must contain $-I$ and the quotient $\rho(\pi_1(N))/\{\pm I\}$ must be cyclic. It follows that in this case we must have $\rho(\pi_1(N)) \cong (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/2\mathbf{Z})$ for some $n \geqslant 0$.

Let S denote the set of all homomorphisms of $\pi_1(N)$ onto groups of the form $(\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/\varepsilon\mathbf{Z})$, where n ranges over the nonnegative integers and ε over $\{1,2\}$. Since $\pi_1(N)$ is finitely generated, S is countable. We can write

$$t^{-1}(\chi) \subset \bigcup_{\phi \in S} A_\phi;$$

here $A_\phi \subset R(\pi_1(N))$ denotes the set of all homomorphisms of the form $\sigma \circ \phi$, where σ ranges over all faithful representations of $\phi(\pi_1(N)) = (\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/\varepsilon\mathbf{Z})$ in $\mathrm{SL}(2, \mathbf{C})$. Such a representation σ must send the generator of $\mathbf{Z}/\varepsilon\mathbf{Z}$ to $-I$ if $\varepsilon = 2$ (and to I if $\varepsilon = 1$). Hence, if we are given $\phi \in S$ and we fix an element $\gamma \in \pi_1(N)$ such that $\phi(\gamma)$ is the standard generator of $(\mathbf{Z}/n\mathbf{Z}) \times \{0\}$, a representation $\rho \in A_\phi$ is uniquely determined by the element $\rho(\gamma)$. If $\rho \in A_\phi \cap t^{-1}(\chi)$ then the trace of $\rho(\gamma)$ must be equal to $\chi(\gamma)$. Since the set of matrices with a given trace is a 2-dimensional subvariety of $\mathrm{SL}(2, \mathbf{C})$, it follows that the set $A_\phi \cap t^{-1}(\chi) \subset R(\pi_1(N))$ has dimension at most 2 for any $\phi \in S$. Thus

$$t^{-1}(\chi) = \bigcup_{\phi \in S} A_\phi \cap t^{-1}(\chi)$$

is a countable union of sets of dimension at most 2. Since we showed that the complex algebraic set $t^{-1}(\chi)$ has dimension 3, we now have a contradiction.

7.3. *The character variety argument completed*

Having finished the proof of Lemma 7.2.1, we return to the main menu. (Hi, my name's Peter, I'm your waiter.) Remember we are trying to find a representation $\rho : \pi_1(M) \to \mathrm{SL}(2, \mathbf{C})$ satisfying (i) and (ii). In view of the lemma (applied to $N = \operatorname{int} M$), it's enough to find a point $\chi \in X_0$ such that every $\rho \in t^{-1}(x)$ satisfies (ii). The simple observation that gets us started in doing this is that if ω is any primitive $2n$th root of unity in $\mathbf{C}$, then any matrix with trace $\omega + \omega^{-1}$ has order $2n$ in $\mathrm{SL}(2, \mathbf{C})$. This is because the matrix

$$\begin{pmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{pmatrix}$$

has the right trace and the right order; and since $n > 1$ we have $\omega + \omega^{-1} \neq \pm 2$, so that any two matrices with trace $\omega + \omega^{-1}$ and determinant 1 are conjugate.

The upshot of all this is that if I_μ takes the value $\omega + \omega^{-1}$ at some point $\chi \in X_0$, then any representation $\rho \in t^{-1}(\chi)$ will satisfy (ii); by the lemma, some $\rho \in t^{-1}(\chi)$ will also satisfy (i). So ρ will give rise to a representation of $|\pi_1(M)\colon\ \mu^n = 1|$ in $\mathrm{PSL}(2, \mathbf{C})$ with a noncyclic image, and in particular it will follow that $|\pi_1(M)\colon\ \mu^n = 1|$ is noncyclic, and hence that $\pi_1(\widetilde{\Sigma})$ is nontrivial.

Now by Corollary 4.5.2, the function I_μ is nonconstant on X_0. Furthermore, as there is a canonical isomorphism $\mathbf{C}(\widehat{X}_0) \to \mathbf{C}(X_0)$, we can extend I_μ to a rational function $\hat{I}_0 : \widehat{X}_0 \to \mathbf{C} \cup \{\infty\}$; since I_0 is nonconstant, $\hat{I}_0$ is surjective. Thus *either* I_μ takes the value $\omega + \omega^{-1}$ at some point of X_0, *or* $\hat{I}_\mu$ takes the value $\omega + \omega^{-1}$ at some point $x \in \widehat{X}_0 - X_0$. In the latter case we can still show that $\pi_1(\widetilde{\Sigma})$ is nontrivial, but by a quite different method.

The idea is, of course, to look at the tree T associated to the ideal point x. Since I_μ takes a finite value at x, it follows from 5.4.2 that μ fixes a vertex of T. Since μ generates the image of the fundamental group of $|\mu| \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap |\mu| = \emptyset$. (In applying the corollary we take $k = 1$ and $B_1 = |\mu|$.) Thus either F has boundary components which are all parallel to $|\mu|$ – i.e., its boundary slope is the meridional slope – or it is closed.

At this point it is easy to reduce the proof of Theorem 7.0.1 to the proof of the following result, which can be proved by arguments due to Meeks–Yau, Gordon–Litherland and Thurston.

THEOREM 7.3.1. *Let K be a prime tame knot in a closed, connected, orientable 3-manifold Σ, such that the exterior M of K is irreducible. Suppose that either M contains a closed essential surface, or the meridian of K is a boundary slope in M. Then for any $n > 1$, the n-fold cyclic branched covering space $\widetilde{\Sigma}$ of Σ branched over K contains a closed, connected essential surface. (Here the term "essential" can be interpreted according to the definition I gave in Section 1, even though we don't know that $\widetilde{M}$ is irreducible.)*

The point is that what I have shown above is that if $\operatorname{int} M$ is hyperbolic, and if we are in the case where the approach to the proof of Theorem 7.0.1 based on the character variety fails, the hypothesis of Theorem 7.3.1 must hold. But the conclusion of 7.3.1 certainly implies the conclusion of 7.0.1, since if $\widetilde{F}$ is essential in $\widetilde{M}$, then according to the definition I gave in Section 1, $\widetilde{F}$ has genus > 0, and $\pi_1(\widetilde{\Sigma})$, which contains an isomorphic copy of $\pi_1(F)$, is therefore nontrivial. If $\operatorname{int} M$ is not hyperbolic we are still OK, because then by Thurston's geometrization theorem, either M contains an essential torus and we can still apply Theorem 7.3.1, or M is Seifert fibered, in which case the proof of Theorem 7.0.1 is an elementary exercise based on the classification of the Seifert fibered spaces.

7.4. *The equivariant loop theorem*

I will give only the briefest hint about the proof of Theorem 7.3.1, since the techniques don't have much to do with the subject of this chapter. If $F \subset M$ is an essential surface which is either closed or has meridional boundary slope, then the pre-image of F in $\widetilde{\Sigma}$, say $\widetilde{F}_0$, is a closed bicollared surface, possibly disconnected but definitely invariant under the action of $\mathbf{Z}/n\mathbf{Z}$ on the cyclic branched covering space $\widetilde{\Sigma}$. Using the primality of K and the irreducibility of M, it's not hard to show that $\widetilde{F}_0$ has a component of genus > 0. If this component is essential, we're happy. If not, the equivariant version of the Dehn-Lemma–Loop-Theorem due to Meeks and Yau [41] allows one to do compressions on F_0 in such a way that the resulting surface F_1 is still invariant under the $\mathbf{Z}/n\mathbf{Z}$ action. The primality and irreducibility again imply that F_1 must have a component of positive genus, and we can repeat the process until we see an essential component appearing. (By a finiteness argument like the one I described in Section 2.4, the process cannot continue indefinitely.)

The big ingredient here is the equivariant Dehn–Loop Theorem. This was first proved by Meeks and Yau using minimal surface techniques. Their proof, which required a lot of hard analysis, was later reinterpreted in a purely combinatorial setting by Edmonds [28], Dunwoody [27] and Jaco and Rubinstein [36].

7.5. *The root-of-unity phenomenon*

In the argument given in Section 7.3, the only ideal points that had to be considered were those where the function I_μ took finite values of the apparently special form $\omega + \omega^{-1}$, where ω is some root of unity. We saw that just by using the finiteness of I_μ at the ideal point we got topological information – that M contains a closed essential surface or that μ is a boundary slope – which was crucial for the proof. It is natural to wonder whether the fact that one obtains a very special kind of finite value, namely $\omega + \omega^{-1}$, where ω is a root of unity, provides additional restrictions on the situation.

From this point of view, the following result, proved in [14], is striking. Suppose we have a compact orientable manifold M bounded by a torus, with int M hyperbolic, and for simplicity suppose that M contains no closed essential surfaces. Suppose that for some nontrivial element α of $\mathrm{im}(\pi_1(\partial M) \to \pi_1(M))$, the function I_α takes a finite value $c \in \mathbf{C}$ at some ideal point x of the curve given by Corollary 4.5.2. Then $c = \omega + \omega^{-1}$ for some root of unity ω.

As a hint about why this should be so, consider a dual surface F to an action on a tree T associated to x. Since by hypothesis F cannot be closed, the arguments of Section 7.3 show that α belongs to the conjugacy class in $\pi_1(M)$ defined by the boundary slope of α. Now consider for a moment the special case in which ∂F is connected. In this case, α is a product of commutators in $\mathrm{im}(\pi_1(F) \to \pi_1(M))$, which by 2.3.1(ii) is a subgroup of the stabilizer of an edge of T. By Property 5.5.3, it follows that $I_\alpha(x) = 2$. This proves the assertion in this case, since we can take $\omega = 1$.

The proof in the general case is a refinement of this argument. If F is a dual surface having the minimal number of boundary components among all surfaces dual to the action, and if some component of F has n boundary components, then ω can be shown to be an nth root of unity.

Nathan Dunfield found a remarkable application of this result in his paper [26]. As I will barely have a chance to mention Dunfield's results in Sections 9 and 10, you will have to look at his paper to appreciate his application of the root-of-unity phenomenon.

7.6. *Extensions of the theorem*

If you examine the proof of Theorem 7.0.1 that I have sketched here, you will see that it really gives a proof of the following stronger result:

THEOREM 7.6.1. *Let K be a nontrivial tame knot in a closed, connected, orientable 3-manifold Σ. Let n be an integer > 1, and let $\widetilde{\Sigma}$ denote the n-fold cyclic regular branched covering of Σ, branched over K. Then either* (i) $\pi_1(\widetilde{\Sigma})$ *has a nontrivial representation in* $\mathrm{PSL}(2, \mathbf{C})$, *or* (ii) $\widetilde{\Sigma}$ *contains an essential surface. Condition* (i) *can be replaced by the stronger condition that the "orbifold group"* $|\pi_1(M)\colon \mu^n = 1|$, *where M is the exterior of K and μ the meridian, has a representation in* $\mathrm{PSL}(2, \mathbf{C})$ *with noncyclic image.*

This "PSL(2, **C**)-version" of the Smith Conjecture is the result which I mentioned earlier as being stronger than the version of the Smith Conjecture proved in [44]. In [20] you will find a proof of an essentially more general result than Theorem 7.6.1, which applies to many noncyclic regular coverings; it is a "PSL(2, **C**)-version" of a generalization of the Smith Conjecture due to Davis and Morgan [24]. Thurston's orbifold theorem, which you may read about in Bonahon's chapter in this Handbook [6], is in turn much stronger than Theorem 7.6.1. (It is also much harder to prove!) In the cases where the "PSL(2, **C**)-versions" give nondegenerate representations in PSL(2, **C**) of the fundamental group of the branched covering – or the "orbifold group" which contains the group of the branched covering as a finite-index subgroup – Thurston's result actually gives a geometric structure on the manifold whose existence implies the existence of such a representation. Furthermore, the geometric structure is invariant under the group of symmetries of the branched covering, and this accounts for the extension of the representation to the "orbifold group": see Bonahon's chapter in this Handbook [6]. Thurston's theorem also applies in more general situations than the other results.

8. Degrees and norms

The material in this section first appeared in [18] and was worked out by Marc Culler and myself.

In Sections 5.6, 6 and 7, very simple properties of algebraic curves were used to prove nontrivial theorems about 3-manifolds. In Sections 5.6 and 6 we used the simple fact that a nonconstant rational function f on a projective algebraic curve C always has at least one pole. In Section 7 we used the essentially equivalent fact that such a function f takes every value in $\mathbf{C} \cup \{\infty\}$ at least once. These facts were applied to the functions $\hat{I}_\gamma$ on a projective completion $\widehat{X_0}$ of a curve X_0 given by Corollary 4.5.2. (Recall that $\widehat{X_0}$ is a curve in $X(\pi_1(N))$, where N is a hyperbolic 3-manifold with one cusp and γ is a nontrivial peripheral element of $\pi_1(N)$.)

These simple facts about a nonconstant rational function f on a projective algebraic curve C can be regarded as consequences of the fact that f is a *branched covering map.* (Coincidentally the same concept came up, one dimension higher, in the last section.) This is an especially natural point of view in the case where C is smooth, and I'll discuss this case first. To say that f is a branched covering map means that there is a finite set $\Psi \subset \mathbf{C} \cup \{\infty\}$ such that $f|C - f^{-1}(\Psi) : C - f^{-1}(\Psi) \to (\mathbf{C} \cup \{\infty\}) - \Psi$ is a covering map. By compactness, it suffices to show that for each point $x \in C$ there exist a neighborhood U of x in C, and homeomorphic identifications of U and $F(V)$ with the unit disk in $\mathbf{C}$, under which $F|U$ becomes the map $z \mapsto z^n$ for some positive integer $n = n_x$. This in turn is true because f is nonconstant and is complex analytic in terms of local coordinates on C and $\mathbf{C} \cup \{\infty\}$. The integer n_x is the degree of the zero of the function $f - c$ if $c = f(x) \in \mathbf{C}$; if f has a pole at x then n_x is the order of the pole. For any $y \in \mathbf{C} \cup \{\infty\}$, we have $y \in \Psi$ if and only if $n_x > 1$ for some $x \in f^{-1}(y)$.

The fact that f is a branched covering map also implies that it has a well-defined *degree.* The degree may be defined as the degree of the covering map $f|C - f^{-1}(\Psi) : C - f^{-1}(\Psi) \to (\mathbf{C} \cup \{\infty\}) - \Psi$. Thus for any $y \in (\mathbf{C} \cup \{\infty\}) - \Psi$, the number of points in $f^{-1}(y)$ is the degree of f. More generally, for any $y \in \mathbf{C} \cup \{\infty\}$, the degree of f is

$$\sum_{x \in f^{-1}(y)} n_x.$$

When y is 0 or ∞ this says that the degree counts the zeros or poles of f with multiplicity, the multiplicity of a zero or pole being its order.

In the general case, where C is not necessarily smooth, we can get a picture of the function f by resolving the singularities of C. In Section 5.3 I mentioned the process of resolving a singularity of a projective curve. If we apply this process successively at all the singular points of C we get a smooth curve, sometimes called the *normalization* of C and denoted C^{ν}, and a generically one–one map $\nu : C^{\nu} \to C$. This curve does not depend on any choices made in constructing it, but is canonically associated with C.

Now if f is a nonconstant rational function on C, and if we set $f^{\nu} = \nu \circ f : C^{\nu} \to \mathbf{C} \cup \{\infty\}$, we can define the degree of f to be the degree of f^{ν}. It is still the case that for any point y lying outside a suitable finite subset of $\mathbf{C} \cup \{\infty\}$, we have $\operatorname{Card} f^{-1}(y) = \deg f$.

If f is a nonconstant rational function on an *affine* curve C, and if $\widehat{C}$ is a projective completion of C whose ideal points are smooth, then f extends to a rational function $\hat{f}$ on $\widehat{C}$, and we can define the degree f to be the degree of $\hat{f}$. The interpretation as the generic number of points in a fiber still works.

The degree of f can also be defined from an algebraic point of view. The map f gives an identification of the function field of C with an extension of the function field of $\mathbf{C}$, which is a field of rational functions in one indeterminate. The degree of f is the degree of this extension. From this point of view there is no distinction between the smooth case and the singular case, or between the affine case and the projective case. For details, see [48].

It turns out that the study of the degree of the functions $\hat{I}_{\gamma} : X_0 \to \mathbf{C}$ has real applications to the study of 3-manifolds. This was first made clear by my joint work with Culler that appeared as Chapter I of [18], and was developed further in some remarkable papers by Boyer and Zhang [8–10]. I will begin the discussion of this degree in the present section. In Sections 9 and 10 I will give some topological applications. All the material in this section is extracted from [18].

8.1. *Degrees of trace functions; defining the norm*

Throughout this section I'll be talking about a hyperbolic 3-manifold N with one cusp; as in the earlier sections I'll let M denote its compact core, I'll choose a curve X_0 with the properties stated in Corollary 4.5.2, and I'll let $\widehat{X}_0$ denote a projective completion of X_0 in which the ideal points are smooth. If we want to calculate the degree of $\hat{I}_{\gamma}$ for a given element $\gamma \in \operatorname{im}(\pi_1(\partial M) \to \pi_1(M))$, it is in a sense simplest to do so by counting the poles of $\hat{I}_{\gamma}$, if only because these all occur at the finitely many ideal points of $\widehat{X}_0$. If $x_1, \ldots, x_n$ are the ideal points, we can write

$$\deg I_{\gamma} = \sum_{i=1}^{n} P_{x_i}(I_{\gamma}), \tag{8.1.1}$$

where $P_{x_i}(f)$ denotes the order of the pole of a function f at x_i in the sense of Section 5.5: thus

$$P_{x_i}(I_{\gamma}) = \max\bigl(0, -v_i(I_{\gamma})\bigr), \tag{8.1.2}$$

where v_i is the valuation of $\mathbf{C}(X_0)$ associated to the ideal point x_i.

In order to understand the nature of the right-hand side of (8.1.2), we consider the tautological representation. Let $*$ be a base point in ∂M, let R_0 be a component of $R(\Gamma)$ that maps onto X_0 (see Section 5.4) and let $\mathcal{P}: \pi_1(M, *) \to \mathrm{SL}(2, \mathbf{C}(R_0))$ denote the tautological representation. This is relevant to understanding the terms in the sum (8.1.2), because by (4.4.1) we have $I_\gamma = \operatorname{trace} \mathcal{P}(\gamma)$. Since the subgroup $\Lambda = \operatorname{im}(\pi_1(\partial M, *) \to \pi_1(M, *))$ is abelian, its image under $\mathcal{P}$ is conjugate in $\mathrm{SL}(2, K)$, where K is some finite extension of $\mathbf{C}(R_0)$, either to a group of diagonal matrices or to a group of matrices of the form $\pm\left(\begin{smallmatrix} 1 & \lambda \\ 0 & 1 \end{smallmatrix}\right)$ with $\lambda \in K$. Actually the second alternative is impossible, because it would make $I_\gamma = \operatorname{trace} \mathcal{P}(\gamma)$ equal to $2 \in \mathbf{C}$, a constant function, for every $\gamma \in \Lambda$, whereas we know from Corollary 4.5.2 that these functions I_γ are all nonconstant. So there is a homomorphism η from Λ to K^*, the multiplicative group of the field K, such that

$$\eta(\gamma) = \begin{pmatrix} \eta(\gamma) & 0 \\ 0 & \eta(\gamma)^{-1} \end{pmatrix}$$

for every $\gamma \in \Lambda$. So for each $\gamma \in \Lambda$ we have

$$I_\gamma = \eta(\gamma) + \eta(\gamma)^{-1}. \tag{8.1.3}$$

To calculate $v_i(I_\gamma)$ from (8.1.3), we first use the extension theorem for valuations, Theorem 5.4.1, to get a valuation w_i of K such that $w_i | \mathbf{C}(V) = d_i v_i$ for some positive integer d_i. Now it's an elementary exercise, starting from the definition of a valuation, to show that if w is a valuation of a field K and f is an element of K, then

$$\max\bigl(0, -w\bigl(f + f^{-1}\bigr)\bigr) = |w(f)|.$$

Putting this together with (8.1.2) and (8.1.3), we get

$$P_{x_i}(I_\gamma) = \frac{1}{d_i} \max\bigl(0, -w_i(I_\gamma)\bigr) = \frac{1}{d_i} \bigl| w_i\bigl(\eta(\gamma)\bigr)\bigr|.$$

To simplify the notation a little, let's set $\ell_i(\gamma) = \frac{1}{d_i} w_i(\eta(\gamma))$, so that $\ell_i : \Lambda \to \mathbf{Z}$ is a homomorphism of abelian groups for $i = 1, \ldots, n$. Then we have

$$P_{x_i}(I_\gamma) = |\ell_i(\gamma)|, \tag{8.1.4}$$

which, combined with (8.1.1), gives

$$\deg I_\gamma = \sum_{i=1}^{n} |\ell_i(\gamma)|. \tag{8.1.5}$$

So the integer-valued function on the rank-two free abelian group Λ that assigns to each $\gamma \in \Lambda$ the degree of I_γ is a function of a very special sort: it is a finite sum of functions, each of which is the absolute value of a homomorphism $\Lambda \to \mathbf{Z}$. To make this look more familiar, it is useful to think of Λ in a slightly different way. Remember that the inclusion homomorphism $\pi_1(\partial M, *) \to \pi_1(M)$ is injective, so that Λ is isomorphic in a canonical way to $\pi_1(\partial M, *)$; since $\pi_1(\partial M, *)$ is abelian, it is in turn canonically isomorphic to $H_1(\partial M, \mathbf{Z})$. Finally, the latter group can be identified with a lattice in the 2-dimensional vector space $V = H_1(\partial M, \mathbf{R})$. So we can identify Λ with this lattice by an isomorphism of groups. When we do this, each of the homomorphisms $\ell_i : \Lambda \to \mathbf{Z}$ can be extended to a linear form $V \to \mathbf{R}$, which I'll still denote ℓ_i. Then, copying the formula (8.1.5), we can define a function $\|\cdot\| : V \to [0, \infty) \subset \mathbf{R}$ by

$$\|u\| = \sum_{i=1}^{n} |\ell_i(\gamma)|, \tag{8.1.6}$$

so that $\|\gamma\| = \deg I_\gamma$ for every $\gamma \in \Lambda$. From the formula (8.1.6) we deduce immediately that

$$\|u_1 + u_2\| \leqslant \|u_1\| + \|u_2\| \tag{8.1.7}$$

for all $u_1, u_2 \in V$, and

$$\|ru\| = |r|\,\|u\| \tag{8.1.8}$$

for all $u \in V,\ r \in \mathbf{R}$.

If V is any vector space, a function $\|\cdot\| : V \to \mathbf{R}$ that satisfies (8.1.7) and (8.1.8) is called a *seminorm*. It's called a *norm* if it also satisfies

$$\|u\| > 0 \tag{8.1.9}$$

for every nonzero vector $u \in V$. Before giving a little context for these definitions, let me give the simple proof that the function $\|\cdot\|$ that I've defined on the 2-dimensional vector space $V = H_1(\partial M, \mathbf{R})$ satisfies (8.1.9) and is therefore actually a norm.

The key point is that (8.1.9) is true if $0 \neq u \in \Lambda$, because then $\|u\| = \deg I_u$, and since I_u is nonconstant according to Corollary 4.5.2, it has a strictly positive degree. Now we certainly can't have $\|u\| = 0$ for *every* $u \in V$, since $\|u\| > 0$ when $u \in \Lambda$; so in the expression (8.1.6) defining $\|\cdot\|$, the linear forms ℓ_i can't all be identically zero. After re-indexing we can assume that ℓ_1 is not identically zero. If for some vector $u_0 \neq 0$ we have $\|u_0\| = 0$, then in particular $\ell_1(u_0) = 0$, so u_0 spans the kernel of ℓ_1. But since, by construction, ℓ_1 maps Λ to $\mathbf{Z}$, the kernel of ℓ_1 is spanned by an element of Λ. So after multiplying u_0 by a nonzero constant we can assume $u_0 \in \Lambda$, and as $\|u_0\| = 0$ we now have a contradiction.

So far we have established the following properties of $\|\cdot\| : V \to \mathbf{R}$:

PROPERTY 8.1.10. *The function* $\|\cdot\| : V \to \mathbf{R}$ *is a norm*;

and

PROPERTY 8.1.11. *For each* $\gamma \in \Lambda$ *we have* $\|\gamma\| = \deg I_\gamma$.

The norm on $V = H_1(\partial M; \mathbf{R})$ that I have described is essentially the same as the one defined in Section 1 of [18]. Actually these two norms differ by a factor of 2; the reason for this will emerge.

I'll sometimes denote this norm $\|\cdot\|$ on $H_1(\partial M)$ by $\|\cdot\|_M$. If you want to justify this notation on strictly logical grounds, you will have to check that the norm depends only on M, and not on the choice of the lift ρ_0 of the discrete faithful representation to $\mathrm{SL}(2, \mathbf{C})$ that was used in defining the norm. It seems to me that this is easy enough to check, but if you don't want to go to the trouble you can just think of $\|\cdot\|_M$ as depending on a choice which is suppressed from the notation.

8.2. *A word about norms*

Before going on, I would like to give a brief discussion of the geometric meaning of norms on a finite-dimensional vector spaces. The most familiar example of a norm on $\mathbf{R}^n$ is of course the Euclidean norm $\|\cdot\|_E$, defined by $\|(x_1, \dots, x_n)\|_E = \sqrt{x_1^2 + \cdots + x_n^2}$. If $\|\cdot\|$ is a norm on a vector space V one can define a metric d on V, generalizing the definition of the Euclidean metric, by setting $d(u, v) = \|u - v\|$. What is interesting about this metric in the finite-dimensional case is not the topology it defines, because it is an elementary fact, sometimes called the "equivalence of norms theorem", that any metric defined by a norm on $\mathbf{R}^n$ gives rise to the same topology as the Euclidean metric.

(In a somewhat sharper form, the equivalence of norms theorem states that for any norm $\|\cdot\|$ there are constants $C, C' > 0$ such that $\|u\| \leqslant C\|u\|_E$ and $\|u\|_E \leqslant C'\|u\|$ for every $u \in \mathbf{R}^n$. To prove the existence of C one writes $u = (x_1, \dots, x_n)$ in the standard basis as $\sum_{i=1}^n x_i e_i$, and uses (8.1.7) and (8.1.8) to conclude that

$$\|u\| \leqslant \sum_{i=1}^{n} |x_i| \, \|e_i\| \leqslant \|u\|_E \sum_{i=1}^{n} \|e_i\|,$$

so that we can take C to be $\sum_{i=1}^n \|e_i\|$. The existence of C implies in particular that $\|\cdot\|$ is continuous in the usual topology of $\mathbf{R}^n$. Since, by (8.1.9), $\|\cdot\|$ is strictly positive on the Euclidean unit sphere S^{n-1}, it now follows that $\|\cdot\|$ takes a minimum value $c > 0$ on S^{n-1}. The proof of the equivalence of norms theorem is now completed by setting $C' = 1/c$ and invoking (8.1.8) again.)

What *is* interesting about a norm (or the associated metric) on a finite-dimensional vector space is the geometric structure to which it gives rise. Specifically, if $\|\cdot\|$ is a norm on a vector space V, then the ball B of radius 1 about the origin, consisting of all $u \in V$ such that $\|u\| \leqslant 1$, is a convex, compact subset of V having 0 as an interior point. (Convexity is immediate from properties (8.1.7) and (8.1.8). In proving that B is compact and that $0 \in \operatorname{int} B$, we may assume that $V = \mathbf{R}^n$; in this case the assertions follow from the equivalence of norms theorem.) It follows from (8.1.9) that B is also *balanced* in the sense that $-u \in B$ whenever $u \in B$. Conversely, if $B \subset V$ is a balanced, convex, compact set with $0 \in \operatorname{int} B$, it is clear that for each $u \in \mathbf{R}^n$ there is a nonnegative real number $r \in [0, \infty)$ such that $u = ru_0$ for some $u_0 \in B$. By compactness there is in fact a least such r, say $r = r_u$. It is a straightforward exercise to check that the function $\|\cdot\| : V \to [0, \infty)$ defined by $\|u\| = r_u$ is a norm, and that this construction is precisely the inverse of the construction that associates to each norm its unit ball. So when V is finite-dimensional we have a canonical bijection between norms on V and balanced, convex, compact sets whose interiors contain the origin. A norm is an appealing algebraic way of encoding the structure of a certain kind of geometric object.

Sometimes I'll find it convenient to look at the ball

$$B_r = \{u \in V\colon \|u\| \leqslant r\}$$

of radius r associated to a norm $\|\cdot\|$, where r is a positive number not necessarily equal to 1. The difference between B_r and B_1 is not a big deal, because (8.1.8) says that $B_r = rB_1 = \{ru\colon u \in B_1\}$. We can also think of B_r as the unit ball for $r\|\cdot\|$, which according to the definition is itself a norm on V.

Another useful fact, which is also easy to prove by the elementary methods I've been talking about, is that the unit sphere of a norm $\|\cdot\|$, i.e., the set of all $u \in V$ with $\|u\| = 1$, is precisely the boundary of its unit ball.

8.3. *Further properties of the norm*

Now I'll discuss some more properties of the norm $\|\cdot\| = \|\cdot\|_M$ on $V = H_1(\partial M)$, where M is the compact core of a one-cusped orientable finite-volume hyperbolic 3-manifold, and the associated balanced convex set $B \subset V$. The first thing to notice is that because expression (8.1.6) that was used to define $\|\cdot\|$ is a finite sum of absolute values of linear forms, the unit sphere ∂B is a polygon, i.e., a finite union of line segments. This may seem almost obvious, but giving a careful proof of it leads to important information. To begin with, we note that of the linear forms ℓ_{x_i} that appear in expression (8.1.6), it may happen that some are identically zero. After reindexing the ℓ_i, if need be, we may assume that there is a natural number $k \leqslant n$ such that

$$\|u\| = \sum_{i=1}^{k} |\ell_i(\gamma)| \tag{8.3.1}$$

for every $u \in V$, and ℓ_i is not identically zero for any $i \leqslant k$. (Actually we must have $k \geqslant 2$, as otherwise $\|\cdot\|$ wouldn't satisfy (8.1.9).) The kernel of ℓ_i, for each $i \leqslant n$, is a line L_i through the origin in the plane V. Each L_i divides the plane into two half-planes, H_i^+ and H_i^-, such that ℓ_i is $\geqslant 0$ on H_i^+ and $\leqslant 0$ on H_i^-.

The lines $L_1, \ldots, L_k$ divide the plane into $2k$ sectors. Let Σ denote any one of these sectors. For each $i \leqslant k$, the sector Σ is contained in either H_i^+ or H_i^-. Hence the ith term $\|\ell_i\|$ in the sum (8.3.1) is identically equal on Σ either to the linear form ℓ_i or to the linear form $-\ell_i$. It follows that $\|\cdot\||\Sigma$ coincides with the restriction to Σ of a function λ_Σ which is a finite sum of linear forms, and is therefore itself a linear form. The intersection of Σ with the unit ball ∂B of $\|\cdot\|$ coincides with the intersection of Σ with the line on which

λ_Σ is equal to 1. This is a line segment. It follows that ∂B is a polygon, as it is the union of the $2k$ line segments $\partial B \cap \Sigma$, where Σ ranges over the sectors.

However, this argument shows more. Since the line segment $\partial B \cap \Sigma$ is the intersection of Σ with a line not passing through 0, the endpoints of this segment must lie on the rays that make up the frontier of Σ. Each of these rays is contained in one of the lines L_i. So every vertex of ∂B lies on one of the L_i. What's neat is that the lines L_i have direct topological meaning in terms of the 3-manifold M. Since the definition of the ℓ_i involved ideal points of the curve $\widehat{X}_0$, you will probably guess that the meaning of the L_i has something to do with essential surfaces; and you will not be wrong.

As I mentioned in Section 5.6, the term "slope" is used to indicate an unoriented isotopy class of simple closed curves in ∂M. These are in bijective correspondence with indivisible elements of $\pi_1(\partial M)$ modulo sign. In this section we are identifying $\pi_1(\partial M)$ with the lattice $\Lambda \subset V$. I'll talk about the "slope of an indivisible element of Λ", so that there are just two indivisible elements with any given slope and they differ by sign. In 5.6 I also pointed out that there is a "boundary slope" associated with each bounded essential surface $F \subset M$. I'll call an element of Λ a *boundary class* if its slope is a boundary slope. Two indivisible elements of Λ which differ by sign span the same 1-dimensional subspace of V. So corresponding to each "slope" there is a line through the origin in the plane V. I'll call a line through the origin a *boundary line* if it corresponds to a boundary slope. According to the theorem of Hatcher's that I talked about in Section 5.6, only finitely many lines in V occur as boundary lines of bounded essential surfaces in M.

PROPOSITION 8.3.2. *Each of the lines $L_1, \dots, L_k$ is the boundary line of some bounded essential surface in M. So each vertex of the polygon ∂B lies on a boundary line.*

To prove this, we recall that for each $i \leqslant k$ the linear form ℓ_i restricts to a nontrivial homomorphism from Λ to $\mathbf{Z}$, so that $\ker \ell_i$ is a direct factor of Λ and is therefore spanned by some indivisible element γ_i, which is the homology class determined up to sign by some simple closed curve $C \subset \partial M$. What we have to prove is that there is a bounded essential surface $F \subset M$ whose boundary components are all isotopic in ∂M to C.

Remember that L_i is the kernel of the linear form ℓ_i, which is defined in terms of an ideal point x_i of X_0. According to Section 5, there is an action of $\pi_1(M)$ on a tree T_i associated to x_i. Since $\gamma_i \in \ker \ell_i$, we find from (8.1.4) that

$$P_{x_i}(I_{\gamma_i}) = |\ell_i(\gamma_i)| = 0.$$

In other words, I_{γ_i} does not have a pole at x_i. It therefore follows from Property 5.4.2 that γ_i fixes a vertex of T_i.

Now we argue as in Section 7. Let $|\gamma_i|$ denote a simple closed curve realizing the slope of γ_i. Since γ_i generates the image of the fundamental group of $|\gamma_i| \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of M on T, such that $F \cap |\gamma_i| = \emptyset$. Thus either F has boundary components which are all parallel to $|\gamma_i|$ – i.e., its boundary class is γ_i – or it is closed.

Of course the conclusion that γ_i is a boundary class is exactly what we want, because it says that L_i is a boundary line. So we need to rule out the possibility that F is closed.

If we assume F is closed, then $\partial M \subset M - F$, so by 2.3.1(i) the subgroup Λ of $\pi_1(M)$ fixes a vertex of T_i. This assertion makes sense even though Λ, as a subgroup of $\pi_1(M)$, is defined only up to conjugation: if a given subgroup fixes some vertex of T, then any conjugate subgroup also fixes some – possibly different – vertex. By Property 5.4.2 and (8.1.4) it then follows that

$$|\ell_i(\gamma)| = P_{x_i}(I_\gamma) = 0$$

for every $\gamma \in \Lambda$, i.e., that ℓ_i is identically zero. But this is false since $i \leqslant k$. The proof of Proposition 8.3.2 is now complete.

Everything I have said up to now about the norm has involved calculating the degrees of the functions I_γ in terms of poles. The degree of a function can also be calculated in terms of its zeros; in the next section I'll show how to get new information about the norm by studying the zeros of certain functions closely related to the I_γ, and comparing this with the information we already have.

9. Applications to Dehn surgery

9.1. *The Cyclic Surgery Theorem*

The first application of the norm $\|\cdot\|_M$ that I talked about in the last section was given in [18], where it was used to prove the Cyclic Surgery Theorem. I will refer you to Boyer's chapter in this volume for the motivation, but I will review the statement of the Cyclic Surgery Theorem here. If M is a compact, orientable 3-manifold whose boundary is a torus, and α is a slope (i.e., an isotopy class of simple closed curves on ∂M), I will denote by $M(\alpha)$ the manifold obtained from M by the Dehn filling with filling slope α. By definition, this means that $M(\alpha)$ is obtained from the disjoint union $M \amalg (S^1 \times D^2)$ by gluing the boundaries of M and $S^1 \times D^2$ via some homeomorphism which maps a simple closed curve representing the slope of α to a curve $\{*\} \times \partial D^2$ for some point $* \in S^1$.

THEOREM 9.1.1 (Culler, Gordon, Luecke and Shalen [18]). *Let M be a compact, orientable 3-manifold whose boundary is a torus. Suppose that M is irreducible but is not a Seifert fibered space. Let α and β be indivisible elements of $\Lambda = \mathrm{im}(\pi_1(\partial M) \to \pi_1(M))$ such that $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ are cyclic. Then the geometric intersection number $\Delta(\alpha, \beta)$ is at most one.*

9.2. *The "no-closed-surface" case*

The proof of the Cyclic Surgery Theorem turns out to be a lot simpler in the special case in which we assume that M contains no closed essential surfaces. I will spend most of this section discussing the proof in this case first. In Section 9.5 I'll come back and say a little about the refinements that one has to make to handle the general case.

So for the rest of this subsection I will be assuming that M contains no closed essential surfaces.

In particular M contains no essential tori; and since the hypothesis of the theorem rules out the Seifert fibered case, it follows from Thurston's geometrization theorem that the interior of M – call it N – has a hyperbolic structure of finite volume. Since ∂M is one torus, N has one cusp. Let's fix a curve X_0 with the properties stated in Corollary 4.5.2. As in Section 8 I'll identify $\Lambda = \operatorname{im}(\pi_1(\partial M) \to \pi_1(M))$ with $H_1(\partial M; \mathbf{Z})$, which I'll think of as a lattice in the 2-dimensional vector space $V = H_1(\partial M; \mathbf{R})$. As in Section 8 we have a norm $\|\cdot\| = \|\cdot\|_M$ defined on V. A key step in the proof of the cyclic surgery theorem in this case turns out to be extracting information about the norm of an element α of Λ from the condition that $\pi_1(M(\alpha))$ is cyclic.

The following simple but arresting lemma, which was discovered by Marc Culler, gives the first indication that the character variety may be useful in studying classes $\alpha \in \Lambda$ for which $M(\alpha)$ is cyclic. It was the starting point for Culler's and my work on this subject.

LEMMA 9.2.1. *Let α be an indivisible element of Λ such that $\pi_1(M(\alpha))$ is cyclic. Let χ be a point of X_0 such that $I_\alpha(\chi) = \pm 2$. Then $I_\gamma(\chi) = \pm 2$ for every $\gamma \in \Lambda$.*

PROOF. By Lemma 7.2.1, applied to $N = \operatorname{int} M$, there is a representation $\rho \in t^{-1}(\chi)$ such that the image of $\rho(\pi_1(M))$ under the natural homomorphism $p : \mathrm{SL}(2, \mathbf{C}) \to \mathrm{PSL}(2, \mathbf{C})$ is noncyclic. If $\rho(\alpha) = \pm I$, then $p \circ \rho : \pi_1(M) \to \mathrm{PSL}(2, \mathbf{C})$ maps α to the identity element of $\mathrm{PSL}(2, \mathbf{C})$ and hence factors through a homomorphism of $|\pi_1(M)\colon \alpha = 1|$ onto the noncyclic group $p(\rho(\pi_1(M)))$. This is impossible, since $|\pi_1(M)\colon \alpha = 1| \cong \pi_1(M(\alpha))$ is noncyclic. So $\rho(\alpha) \neq \pm I$. On the other hand, we have $\operatorname{trace} \rho(\alpha) = I_\alpha(\chi) = \pm 2$. So after composing ρ with a suitable inner automorphism of $\mathrm{SL}(2, \mathbf{C})$ we can assume that

$$\rho(\alpha) = \pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(Of course a conjugation doesn't change the character of ρ.)

Now for any $\gamma \in \Lambda$ the matrix $\rho(\gamma)$ commutes with

$$\rho(\alpha) = \pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

since Λ is abelian. So $\rho(\gamma)$ must have the form $\pm\left(\begin{smallmatrix} 1 & \lambda \\ 0 & 1 \end{smallmatrix}\right)$. In particular $I_\gamma(\chi) = \operatorname{trace} \rho(\gamma) = \pm 2$. This proves the lemma. Is that cool, or what? □

In thinking about the consequences of Lemma 9.2.1, it's nice to think in terms of the functions $f_\gamma = I_\gamma^2 - 4 : X_0 \to \mathbf{C}$ which are defined for all $\gamma \in \Lambda$. It's very easy to see that squaring a function doubles its degree, and that adding a constant doesn't change the degree, so in terms of the norm $\|\cdot\|_M$ defined in Section 8 we have

$$\deg f_\gamma = 2 \deg I_\gamma = 2\|\gamma\| \tag{9.2.2}$$

for every $\gamma \in \Lambda$. Now Lemma 9.2.1 says that if $\pi_1(M(\alpha))$ is cyclic, then at every point of X_0 where f_α takes the value 0, all the functions f_γ for $\gamma \in \Lambda$ take the value 0.

Recall from Corollary 4.5.2 that if γ is any nontrivial element of Λ then I_γ is nonconstant on X_0; hence so is f_γ. If we allow ourselves to think in very fuzzy terms for a moment, we can think of the degree of f_γ as counting the points where f_γ takes the value 0, and Lemma 9.2.1 says that f_γ, for an arbitrary $\gamma \in \Lambda$, takes the value 0 wherever f_α does; this suggests the

FUZZY IDEA 9.2.3. *Maybe we should expect to have* $\deg f_\alpha \leqslant \deg f_\gamma$, *and hence* $\|\alpha\| \leqslant \|\gamma\|$, *for every* $\gamma \in \Lambda$, *if* $\pi_1(M(\alpha))$ *is cyclic. In other words, when* $\pi_1(M(\alpha))$ *is cyclic, maybe the the norm of* α *should be minimal among all norms of nontrivial elements of the lattice* Λ.

However, we have not come even close to proving this. This is because calculating the degree of a function in terms of its zeros makes sense only when the domain is a smooth projective curve, and even in this case the orders of the zeros must be taken into account. The curve X_0 is not projective, it need not be smooth, and we have ignored orders.

Of these three issues, the last two turn out to be essentially technical. First of all, the proof of Lemma 9.2.1 can be souped up, by an argument involving the tautological representation and valuations of function fields, to show that if χ is a smooth point of X_0 where f_α has a zero of a certain order, then f_γ has a zero of at least the same order. To put this in symbols, let's write $Z_x(f)$ to mean the order of zero of a function f at a smooth point of a curve, and to mean 0 if f does not have a zero at x: thus $Z_x(f) = \max(0, v_x(f))$, where v_x is the valuation of the function field associated to the smooth point x. Then the souped-up version of Lemma 9.2.1 says that if α and γ are elements of Λ with α indivisible and $\pi_1(M(\alpha))$ cyclic, then $Z_\chi(f_\alpha) \leqslant Z_\chi(f_\gamma)$ for every smooth point χ of X_0.

There's also a version of this that works for nonsmooth points of X_0. Let's set $X_0^\nu = \nu^{-1}(X_0) \subset \widehat{X}_0^\nu$. The following result is the ultimate version of Lemma 9.2.1.

THEOREM 9.2.4. *Let* α *be an indivisible element of* Λ *such that* $\pi_1(M(\alpha))$ *is cyclic. Then for every point* x *of* X_0^ν *we have*

$$Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu).$$

Again, the most important thing to say about the proof of this theorem is that it's essentially an elaborate technical refinement of the simple proof of Lemma 9.2.1. As amusing as the algebra involved is, I will fight back the impulse to give any of the argument, but I will make one comment to help get you into the right mood for reading the details in [18]. Although on the face of it X_0^ν would appear to be only a subset of a projective curve, it turns out to have the structure of an affine curve, the *affine normalization* of X_0. From the algebraic point of view, the coordinate ring $\mathbf{C}[X^\nu]$ is the *integral closure* of $\mathbf{C}[X]$.

Theorem 9.2.4 deals with two of the issues I mentioned above in connection the proposed "proof" of the Fuzzy Idea 9.2.3. The remaining issue concerns ideal points, and is different in nature. In fact, it should be clear by now that anything related to ideal points of X_0 has something to do with the topology of M, specifically with the essential surfaces in M.

The natural context for all this is provided by the normalized projective completion $\widehat{X}_0^\nu$. According to the discussion in the introduction to Section 8, we have

$$\deg f = \deg \hat{f}^\nu = \sum_{x \in \widehat{X}_0^\nu} Z_x(\hat{f}^\nu). \tag{9.2.5}$$

Of course this last expression is really a finite sum since $Z_x(\hat{f}^\nu) = 0$ for all but finitely many points $x \in \widehat{X}_0^\nu$.

Now let α be an indivisible element of Λ such that $\pi_1(M(\alpha))$ is cyclic. According to (9.2.5), the inequality proposed in 9.2.3 will hold if $Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu)$ for every $x \in \widehat{X}_0^\nu$; and according to Theorem 9.2.4, this is true whenever $x \in X_0^\nu$.

Now suppose that $x \in \widehat{X}_0^\nu - X_0^\nu$. We would like to know that $Z_x(f_\alpha^\nu) \leqslant Z_x(f_\gamma^\nu)$ in this case. This is not even an issue unless $Z_x(f_\alpha^\nu) \neq 0$, so that $I_\alpha(\nu(x)) = \pm 2$. Suppose this happens, and let's look at the tree T associated to the ideal point $\nu(x)$ of $\widehat{X}_0$. Since I_α takes the finite value 2 at x it follows from 5 that α fixes a vertex of T. Let C be a simple closed curve in ∂M realizing the slope of α. Since α generates the image of the fundamental group of $C \subset \partial M$ in $\pi_1(M)$, it follows from Corollary 6.0.1 that there is an essential surface $F \subset M$, dual to the action of $\pi_1(M)$ on T, such that $F \cap C = \emptyset$. Thus either F has boundary components which are all parallel to C – i.e., its boundary slope is the slope of α – or it is closed.

We are excluding from the discussion the case where M contains closed essential surfaces, so we need only worry about the case where F is bounded and has α as a boundary class. In this case, where α is a boundary class of M, we need to resort to some serious topology. The relevant argument here is due to Gordon. If F is any essential surface in M having α as a boundary class, we can use F to construct a closed bicollared surface $F(\alpha)$ in the manifold $M(\alpha)$. After all, $M(\alpha)$ is constructed from M by attaching a solid torus K in which α is a meridian; since the components of ∂F are meridians in K, we can attach disjoint disks in K to the boundary components of F, and this gives the surface $F(\alpha)$. Let me suspend the conventions of this subsection for a moment so that I can state the following result with all relevant hypotheses so as to make its self-contained nature clear.

THEOREM 9.2.6 (Gordon). *Let M^3 be compact, connected, orientable and irreducible, with ∂M a torus. Suppose that M contains no closed essential surfaces. Let α be a boundary class for M. Among all essential surfaces in M having α as a boundary class, let us choose one, say F_0, whose number of boundary components is minimal. Then either* (i) *F has strictly positive genus and $F(\alpha)$ is an essential surface in $M(\alpha)$, or* (ii) *F has genus* 0, *so that $F(\alpha)$ is a 2-sphere; and either F is a fiber in a fibration of M over S^1, or $F(\alpha)$ decomposes $M(\alpha)$ as the connected sum of two nontrivial lens spaces.*

A lens space is *nontrivial* if it's not a 3-sphere. To say that $F(\alpha)$ decomposes $M(\alpha)$ as the connected sum of two nontrivial lens spaces means that the closure of each component of $M(\alpha) - F(\alpha)$ is homeomorphic to the manifold obtained from a nontrivial lens space by removing the interior of a 3-ball with bicollared boundary. Such a ball is unique up to ambient isotopy in $M(\alpha)$; see [35].

Note that if (i) holds then $\pi_1(M(\alpha))$ contains an isomorphic copy of $\pi_1(F(\alpha))$, where $F(\alpha)$ is a closed, orientable surface of positive genus; and that if (ii) holds, $\pi_1(M(\alpha))$ is a free product of two nontrivial cyclic groups. So:

COROLLARY 9.2.7. *Under the hypotheses of Theorem* 9.2.6, $\pi_1(M(\alpha))$ *is not cyclic.*

Resuming the conventions of this subsection, we can now go back to considering a class $\alpha \in \Lambda$ such that $\pi_1(M(\alpha))$ is cyclic. If I_α takes a finite value at some ideal point of X_0 then we have seen that, under the assumption that M contains no closed essential surfaces, α must be a boundary class; and this contradicts Corollary 9.2.7. So I_α must have a pole at every ideal point, which means that the inequality of Theorem 9.2.4 holds even when $x \in \widehat{X}_0^\nu - X_0^\nu$, since in that case the left-hand side is zero. So we can sum the inequality over all $x \in \widehat{X}_0^\nu$, and according to (9.2.5) we get

$$\deg f_\alpha \leqslant \deg f_\gamma$$

whenever α and γ are elements of Λ such that α is indivisible and $\pi_1(M(\alpha))$ is cyclic. Thus the "fuzzy idea" 9.2.3 turns out to be entirely justified in the case where M contains no closed essential surfaces, and we may state it as a

THEOREM 9.2.8. *If* $\pi_1(M(\alpha))$ *is cyclic then* $\|\alpha\|$ *is minimal among all norms of nonzero elements of* Λ. (*Let me emphasize that this theorem depends on the blanket assumption made in this section that M contains no closed essential surfaces.*)

In order to make the transition between Theorem 9.2.8 and the conclusion of the Cyclic Surgery Theorem, we set $m = \min_{0 \neq \gamma \in \Lambda} \|\gamma\|$, so that m is a natural number and any $\alpha \in \Lambda$ with $\pi_1(M(\alpha))$ cyclic has norm m. The argument is based on considering the ball B_m of radius m for the norm $\|\cdot\| = \|\cdot\|_M$. From what I said in Sections 8.2 and 8.3 it follows that the unit ball B of $\|\cdot\|$ is a compact, convex, balanced set bounded by a polygon, and hence so is $B_m = mB$. We can paraphrase Theorem 9.2.8 by saying that int B_m contains no points of Λ except 0.

It happens that number-theorists have long been interested in properties of convex, balanced polygons whose interiors contain no points of a give lattice. In the next subsection I will state and prove a theorem on this subject, which is one of the simplest results in Minkowski's "geometry of numbers", and illustrate how powerful it is in number theory. Then, in Section 9.4, I will show how to use Minkowski's theorem to finish the proof of the Cyclic Surgery Theorem in the case where there is no closed essential surface.

9.3. *A little geometry of numbers: A theorem of Minkowski's*

Suppose that V is an n-dimensional vector space. The volume element (Lebesgue measure) on V, which we can think of as a function that assigns a nonnegative real number to every bounded, measurable set in V, is well-defined up to multiplication by a positive constant. A convex set is always measurable, so it has a well-defined volume once we have fixed a volume element. If Λ is a lattice in V, i.e., a discrete, cocompact subgroup of the additive group of V, then the quotient V/Λ, which is topologically an n-torus, inherits a volume element when we fix a volume element on V; the volume of the whole torus V/Λ is then called the covolume of Λ. Of course, multiplying the volume element by a positive constant has the effect of multiplying the covolume by the same constant. Hence the conclusion of the following theorem is independent of the choice of a volume element.

THEOREM 9.3.1 (Minkowski). *Let Λ be a lattice in an n-dimensional vector space V. Let B be a compact, convex, balanced subset of V such that* $\operatorname{int} B$ *contains no point of Λ except* 0. *Then*

$$\operatorname{vol} B \leqslant 2^n \operatorname{covol} \Lambda.$$

PROOF. We may take the volume element to be normalized so that $\operatorname{covol} \Lambda = 1$. It follows from convexity that $\operatorname{int} B$ has the same volume as B. The volume element on V induces a volume element on $V/2\Lambda$ as well as on V/Λ. Since V/Λ has volume 1, the volume of $V/2\Lambda$ is 2^n. So it's enough to show that the natural map $p: V \to V/2\Lambda$ maps $\operatorname{int} B$ injectively.

Let x and y be points of B with $p(x) = p(y)$; then $x - y \in 2\Lambda$, i.e., $\frac{x-y}{2} \in \Lambda$. But since $\operatorname{int} B$ is convex and balanced, and contains x and y, we have $\frac{x-y}{2} \in B$. The hypothesis concerning B now implies that $\frac{x-y}{2} = 0$, so $x = y$. This is all there is to the proof of the theorem.

(This is logically equivalent to the proof you will find in books on the subject, but the use of the torus $V/2\Lambda$ makes it more intuitive. John Morgan pointed out this version of the argument to me several years ago.) □

I will illustrate the use of Minkowski's theorem in number theory by showing how it can be used to prove Lagrange's famous theorem that every positive integer is the sum of four squares. I've adapted this from [51]. One way of paraphrasing the statement that a given n is a sum $a^2 + b^2 + c^2 + d^2$ of four squares of integers is to say that n is the squared absolute value $|h|^2$ of a (Hamiltonian) quaternion $h = a + bi + cj + dk$ with integer "coordinates". Since we have $|hk| = |h|\,|k|$ for any quaternions h and k, the property of being a sum of four squares is preserved under the formation of products; so it's enough to prove that every prime p is the sum of four squares. The case $p = 2$ is clear, so let's take p odd.

Having made this reduction, we now forget about quaternions, and reinterpret our goal more naïvely as showing that p is the square of the Euclidean norm $\|v\|$ of a point $v = (a, b, c, d)$ in the standard lattice $\mathbf{Z}^4 \subset \mathbf{R}^4$. The approach is to construct a subgroup Λ of $\mathbf{Z}^4$ which has index p^2, and is therefore a lattice of covolume p^2 in terms of the Euclidean volume element, and such that for every $v \in \Lambda$ we have $\|v\|^2 \equiv 0 \pmod p$. If we have such a Λ, and if B denotes the ball of radius $\sqrt{2p}$ about the origin, then by elementary calculus we find that

$$\operatorname{vol} B = 2\pi^2 p^2 > 16p^2 = 2^4 \operatorname{covol} \Lambda,$$

so that by Minkowski's theorem, there is a nonzero element v of $\Lambda \cap \operatorname{int} B$. Then $0 < \|v\|^2 < 2p$, but the property we're assuming for Λ says that $\|v\|^2$ is divisible by p; so we must have $\|v\|^2 = p$, which is what we need.

The best way to find a Λ with the right properties is to think of $\mathbf{R}^4$ in yet a third way, as the complex vector space $\mathbf{C}^2$, in which case the lattice $\mathbf{Z}^4$ becomes $\mathcal{O}^2$, where $\mathcal{O} = \mathbf{Z}[\mathrm{i}]$ is the ring of Gaussian integers, consisting of all complex numbers $a + \mathrm{i}b$ with $a, b \in \mathbf{Z}$. Given any $s \in \mathcal{O}$, we can define a homomorphism of additive groups $H : \mathcal{O}^2 \to \mathcal{O}$ by $H_s(z, w) = z - sw$. Then $\Lambda = \Lambda_s = H_s^{-1}(p\mathcal{O})$ is a subgroup of index p^2 in $\mathcal{O}^2$, and for any $(z, w) \in \Lambda$ we have $|z|^2 \equiv |s|^2|w|^2 \pmod p$. Hence if we can choose $s \in \mathcal{O}$ so that $|s|^2 \equiv -1 \pmod p$, it will follow that $\|(z, w)\|^2 = |z|^2 + |w|^2 \equiv 0 \pmod p$ for any $(z, w) \in \Lambda$, as required.

But the existence of such an s is easy. The homomorphism $x \to x^2$ from the multiplicative group $(\mathbf{Z}/p\mathbf{Z})^*$ to itself has kernel $\{\pm 1\}$ and hence has image of order $(p-1)/2$, so the set S of squares in $\mathbf{Z}/p\mathbf{Z}$ has cardinality $(p+1)/2$. It follows that the set $T \subset \mathbf{Z}/p\mathbf{Z}$ consisting of elements of the form $1 - x^2$ also has $(p+1)/2$ elements, and the paucity of pigeon holes forces S and T to intersect. So there exist integers u and v such that $-1 - u^2 \equiv v^2 \pmod p$, and $s = u + \mathrm{i}v$ then satisfies $|s|^2 \equiv -1 \pmod p$.

9.4. *The "no-closed-surface" case, concluded*

To complete the proof of the Cyclic Surgery Theorem in the case where M contains no closed essential surfaces, we apply Minkowski's theorem to the set B_m in our 2-dimensional vector space $V = H_1(M, \mathbf{R})$. Let's normalize the volume element on V in such a way that our lattice $\Lambda = H_1(M, \mathbf{Z})$ has co-area 1. The relevance of the theorem comes from the observation that, under this normalization, if γ and γ' are two elements of Λ, then the area of the parallelogram with vertices $\pm\gamma, \pm\gamma'$ is just $2\Delta(\gamma, \gamma')$.

Suppose that α and β are indivisible elements of Λ with $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ cyclic. By Theorem 9.2.8 we have $\|\alpha\| = \|\beta\| = m$, so that $\alpha, \beta \in \partial B_m$. The parallelogram P with vertices $\pm\alpha, \pm\beta$ is therefore contained in B_m, so that

$$\Delta(\alpha, \beta) = \frac{1}{2} \operatorname{area} P \leqslant \frac{1}{2} \operatorname{area} B \leqslant 2, \tag{9.4.1}$$

where in the last step we have used Theorem 9.3.1 to conclude that area $B \leqslant 4$.

To complete the proof of the Cyclic Surgery Theorem in this case we need only rule out the possibility that all the inequalities in (9.4.1) are equalities. If this happens, then in particular α and β are vertices of B_m. Now according to Proposition 8.3.2, each of the vertices of B, and hence of B_m, lies on a boundary line defined by some bounded essential surface in M. So in this situation we conclude that α and β are boundary classes. However, in the situation of this section, knowing that even *one* of the classes α and β is a boundary class is enough to give a contradiction. This follows from Corollary 9.2.7, since $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ have both been assumed cyclic, and M contains no essential surfaces.

And that is how the Cyclic Surgery Theorem is proved in the special case I've been talking about.

9.5. *The case where there are closed surfaces*

In the general case, the proof of the Cyclic Surgery Theorem breaks up into three cases. For two of the cases – the case in which either α or β is a boundary slope, and the case in which $\operatorname{int} M$ is not hyperbolic – I will refer you to Boyer's chapter in this volume. In the case that neither α nor β is a boundary slope, the proof is a refinement of the proof that I gave in the special case where M contains no closed essential surfaces. The main step is to prove the following result, which is mostly a refinement of the proof of Theorem 9.2.8.

THEOREM 9.5.1. *Let M be a manifold with a single torus boundary such that* $\operatorname{int} M$ *has a hyperbolic structure of finite volume, and let $\|\cdot\|$ be the norm on $\Lambda = H_1(\partial M)$ defined in* Section 8. *Let α be an indivisible element of Λ which is not a boundary class, and suppose that $\pi_1(M(\alpha))$ is cyclic. Then $\|\alpha\|$ is minimal among all norms of nonzero elements of Λ.*

The basic strategy used in the proof is the same as in the proof of Theorem 9.2.8: one shows that for any point x of the curve $\widehat{X}^v$ where $f(\alpha)$ has a zero, f_γ has a zero of at least the same order for each nontrivial element γ of Λ. When $x \in X^v \subset \widehat{X}^v$ this is proved in exactly the same way as above. If x projects to an ideal point of $\widehat{X}$, the above arguments give an action of $\pi_1(M)$ on a tree T under which α fixes a vertex. As before, we can associate an essential surface F with this action, and we can take it to be disjoint from a simple closed curve realizing the slope of α. If F had a nonempty boundary, its boundary slope would be α, and we would have a contradiction. The difficulty is that F may now be closed. Section I.6 of [18] is devoted to the proof that in this situation, if $Z_x(f_\alpha) > Z_x(f_\gamma)$, then we can replace a given closed surface F dual to the action of $\pi_1(M)$ on T by a new dual surface which has nonempty boundary, and has α as a boundary class. The starting point is the observation that since $\pi_1(M)$ is cyclic, there must be a compressing disk for F in M. See [18] for more.

9.6. *Other applications to surgery*

In [10], Boyer and Zhang proved the following analogue of the Cyclic Surgery Theorem:

FINITE SURGERY THEOREM (Boyer–Zhang). *Let M be a compact, orientable* 3*-manifold whose boundary is a torus. Suppose (for simplicity) that* $\operatorname{int} M$ *has a hyperbolic metric of finite volume. Let α and β be indivisible elements of Λ such that $\pi_1(M(\alpha))$ and $\pi_1(M(\beta))$ are finite. Then the geometric intersection number $\Delta(\alpha, \beta)$ is at most three. Furthermore, up to sign there are at most five indivisible elements α of Λ such that $\pi_1(M(\alpha))$ is finite.*

The bounds of three and five are both sharp. For more discussion of the context of the statement, see Boyer's chapter in this Handbook [7].

A key step in the proof of the Finite Surgery Theorem is the following surprising analogue of Theorem 9.5.1, which was proved in [8]; I will have to refer you to [8] for an account of the very remarkable new ideas that enter into the proof.

THEOREM 9.6.1. *Let M be a manifold with a single torus boundary such that* $\operatorname{int} M$ *has a hyperbolic structure of finite volume, and let $\|\cdot\|$ be the norm on $\Lambda = H_1(\partial M)$ defined in Section* 8. *Let α be an indivisible element of Λ which is not a boundary class, and suppose that $\pi_1(M(\alpha))$ is finite. Then*

$$\|\alpha\| \leqslant \max\{2m, m+8\}, \quad \text{where } m = \min_{0 \neq \gamma \in \Lambda} \|\gamma\|.$$

Using Theorem 9.6.1 and the kinds of arguments that I've discussed earlier in this chapter, Boyer and Zhang were able to show that the bound of 3 on the geometric intersection number asserted in the Finite Surgery Theorem holds unless the unit ball for the norm $\|\cdot\| = \|\cdot\|_M$ is of a special type. The same argument works if one replaces the usual norm $\|\cdot\|_M$ by a slightly different norm $\|\cdot\|'_M$; the definition of $\|\cdot\|'_M$ is just like that of $\|\cdot\|_M$, except that in place of the usual curve X_0 one uses the possibly reducible curve X_1 obtained by saturating X_0 under the action of the Galois group of $\mathbf{C}$ over $\mathbf{Q}$. The completion of the proof of the Finite Surgery Theorem given in [10] essentially involves showing that polygons of these exceptional types do not arise as unit balls of norms $\|\cdot\|'_M$ for 3-manifolds M whose interiors have 1-cusped hyperbolic structures. In order to do this, Boyer and Zhang needed to interpret the polygons arising from 3-manifolds from a new point of view, based on the theory of the so-called A-polynomial, which was developed in [14] and [16], among other papers.

If M is a compact 3-manifold with torus boundary, there is a natural map r from the character variety $X(\pi_1(M))$ to $X(\pi_1(\partial M))$: the image under r of a character of $\pi_1(M)$ is its precomposition with the inclusion homomorphism $\pi_1(\partial M) \to \pi_1(M)$. Basically the A-polynomial – or the slight variant of it used in [8], which I'll call the A'-polynomial – gives information about the image $r(X_1)$ in the case where $\operatorname{int} M$ has a 1-cusped hyperbolic structure and X_1 is the Galois-saturated curve I've just described. It turns out that there is a curve $Y \subset \mathbf{C}^2$ which admits a canonical degree-two rational map to $r(X_1)$. The A'-polynomial of M is a canonically defined two-variable integer polynomial whose locus of zeros is Y. (The A-polynomial is defined similarly except that in place of X_1 one uses a curve X_2 that's possibly still bigger than X_1 in the sense that it may have still more irreducible components.)

Boyer and Zhang re-interpreted the unit ball of the norm $\|\cdot\|_M$ as a "geometric dual" to the so-called Newton polygon of the A'-polynomial, which is another convex plane polygon obtained from the curve A' by an algebro-geometric construction. Then they adapted to the A'-polynomial properties of the A-polynomial established in [14], and the more surprising properties established by Cooper and Long in [16], to deduce restrictions on the unit ball of $\|\cdot\|_M$ which rule out the exceptional polygons that arise in the proof of the Finite Surgery Theorem. For a survey of the relevant material on the A-polynomial, see Cooper and Long's paper [17].

In [26], Dunfield established a fundamental property of the map r, namely that $r|X_0 : X_0 \to r(X_0)$ is a birational map, which is to say that it has degree one. His ingenious proof, besides using the result from [14] about the exactness of Hodgson's volume form, uses deeper properties of hyperbolic manifolds than had previously been brought to bear on the theory that I've been describing in this chapter: Thurston's Dehn surgery theorem and, crucially, the volume rigidity theorem of Goldman, Gromov and Thurston.

Dunfield gave elegant applications of his birationality theorem to Dehn surgery and to related topics. One of his results on Dehn surgery states that if a hyperbolic knot in S^3 is small, i.e., if its exterior contains no closed essential surface, and if it admits a nontrivial cyclic surgery, then it admits a nonintegral boundary slope. (For the terminology I'm using here, it's best to see Boyer's chapter. I'll talk about another of Dunfield's consequences of his birationality theorem in the next section.)

10. Boundary slopes and genera of essential surfaces in knot exteriors: The Neuwirth Conjecture revisited and the Poincaré Conjecture approached

According to Theorem 5.6.2, if a tame knot in any closed orientable 3-manifold Σ satisfies some mild conditions then it has at least two boundary slopes. In the case where $\pi_1(\Sigma)$ is cyclic (for example if $\Sigma = S^3$) one gets strictly stronger results. Some of these are best understood in terms of explicitly identifying slopes as elements of $\mathbf{Q} \cup \{\infty\}$, by the formalism explained in Boyer's chapter in this volume. Recall that if $\mu \in \Lambda = H_1(\partial M)$ denotes the meridian of M, and if we choose an element $\lambda \in \Lambda$ such that $\{\mu, \lambda\}$ is a basis of λ, then we get a bijection between slopes and elements of $\mathbf{Q} \cup \{\infty\}$ by letting the slope of $a\mu + b\lambda$ correspond to a/b. As a/b has at least as much of a right to be called a slope as the corresponding unoriented isotopy class, I will allow myself to blur the distinction between the two when λ has been chosen, or – as in what we're about to see – when the choice doesn't matter.

Changing the choice of λ has the effect of subjecting all slopes to an integer translation, possibly followed by a change of sign. In particular, the absolute value of the difference between two slopes does not depend on the choice of λ. Indeed, a straightforward calculation shows that the difference between the slopes of two indivisible elements $\alpha, \beta \in \lambda$ is given in invariant form by the expression

$$\frac{\Delta(\alpha, \beta)}{\Delta(\alpha, \mu)\Delta(\beta, \mu)},$$

where Δ denotes the geometric intersection number of α and β as in Theorem 9.1.1.

The slope of the meridian is always ∞. When ∞ is not a boundary slope, the set of boundary slopes is a subset of $\mathbf{Q}$ which is finite by Hatcher's theorem [33] and is well-defined modulo integer translation and change of sign. In particular, the *diameter* of the set of boundary slopes – the difference between its greatest and least elements – is well defined.

10.1. *A lower bound for the diameter of the boundary slopes*

The following result is proved in [22].

THEOREM 10.1.1. *Let K be a nontrivial tame knot in a homotopy 3-sphere. Suppose that ∞ is not a boundary slope of K. Then the diameter of the set of boundary slopes of K is $\geqslant 2$.*

An immediate corollary to this theorem is that if ∞ is not a boundary slope, there is always a boundary slope of absolute value at least 1. This is a small step in the direction of the conjecture that there is always a boundary slope which is a nonzero integer. You should think of this conjecture as a stronger form of the conjecture of Neuwirth's that I sketched the proof of in Section 6; in fact it is closely related to some of the stronger versions stated by Neuwirth himself.

Theorem 10.1.1 is deduced in [22], by a relatively routine argument, from the following more general result, Theorem 10.1.2. A *cable knot* is a knot that lies on the boundary of some tame solid torus and has intersection number at least 2 with a meridian disk for that solid torus.

THEOREM 10.1.2. *Let Σ be a closed, connected, orientable 3-manifold such that $\pi_1(\Sigma)$ is cyclic. Let K be a nontrivial tame knot in Σ which is not a cable knot. Suppose that ∞ is not a boundary slope of K. Then the diameter of the set of boundary slopes of K is $\geqslant 2$.*

This result, like Theorem 5.6.2, is technically easier to prove in the case where $\Sigma - K$ is hyperbolic. The proof in the hyperbolic case, which I will give here, turns out to be a remarkably simple application of Theorem 9.5.1. The proof in the general case involves souping up the argument that works in the hyperbolic case in rather the same way that the arguments of Section 5.6 were souped up in Section 6, although the details – for which I will refer you to [22] – are more involved.

In proving Theorem 10.1.2 in the hyperbolic case, it's nice to think in terms of the invariant description of the difference of two slopes that I gave above. From this point of view, what we have to prove is that there exist two boundary classes α and β such that

$$\frac{\Delta(\alpha,\beta)}{\Delta(\alpha,\mu)\Delta(\beta,\mu)} \geqslant 2. \tag{10.1.3}$$

As in Section 9, let's set $m = \min_{0\neq\gamma\in\Lambda} \|\gamma\|_M$. If μ denotes the meridian of K (in the sense of Boyer's chapter, for example), then the "Dehn-filled" manifold $M(\mu)$ is just Σ, which has cyclic fundamental group, so that $\|\mu\| = m$ by Theorem 9.5.1. Thus if B_m denotes the ball of radius m for the norm $\|\cdot\| = \|\cdot\|_M$, we have $\mu \in \partial B_m$.

Now recall from Section 8.3 that the unit ball B of $\|\cdot\|$ is a compact, convex, balanced set bounded by a polygon, and that each vertex of the polygon lies on a boundary line; hence the same is true of $B_m = mB$. Let e denote an edge of the polygon ∂B_m containing μ. The endpoints of e are vertices α_0 and β_0 of ∂B_m, and are therefore positive multiples of boundary classes α and β. I'll complete the proof by showing that inequality (10.1.3) holds with these choices of α and β.

The geometric intersection number $\Delta(\cdot,\cdot)$ is the absolute value of an alternating integer-valued bilinear pairing on the lattice $\Lambda = H_1(\partial M; \mathbf{Z})$. Let's extend the latter pairing to an alternating real-valued bilinear pairing on the vector space $V = H_1(\partial M; \mathbf{R})$; I'll write $\Delta(\cdot,\cdot)$ for the absolute value of this extended pairing as well. If we write $\alpha = a\|\alpha_0\|$

and $\beta = b\|\beta_0\|$, with $a, b > 0$, then we have $\Delta(\alpha, \beta) = ab\Delta(\alpha_0, \beta_0)$, whereas $\Delta(\alpha, \mu) = a\Delta(\alpha_0, \mu)$ and $\Delta(\beta, \mu) = b\Delta(\beta_0, \mu)$. Hence

$$\frac{\Delta(\alpha, \beta)}{\Delta(\alpha, \mu)\Delta(\beta, \mu)} = \frac{\Delta(\alpha_0, \beta_0)}{\Delta(\alpha_0, \mu)\Delta(\beta_0, \mu)}. \tag{10.1.4}$$

Since μ is on the segment e with endpoints α_0, β_0, we can write $\mu = t\alpha_0 + (1-t)\beta_0$ for some $t \in [0, 1]$. So we have

$$\Delta(\alpha_0, \mu) = (1-t)\Delta(\alpha_0, \beta_0) \quad \text{and} \quad \Delta(\beta_0, \mu) = t\Delta(\alpha_0, \beta_0).$$

Combining this with equality (10.1.4), we get

$$\frac{\Delta(\alpha, \beta)}{\Delta(\alpha, \mu)\Delta(\beta, \mu)} = \frac{1}{t(1-t)\Delta(\alpha_0, \beta_0)}. \tag{10.1.5}$$

As I pointed out in Section 9, it follows from Theorem 9.3.1 that B_m has area at most 4. Now we reason as in the proof of Theorem 9.1.1, but with α_0 and β_0 playing the roles of α and β in that argument. The parallelogram P with vertices $\pm\alpha, \pm\beta$ is therefore contained in B_m, and we have

$$\Delta(\alpha_0, \beta_0) = \frac{1}{2} \operatorname{area} P \leqslant \frac{1}{2} \operatorname{area} B_m \leqslant 2.$$

Combining this inequality with equality (10.1.5) we find that

$$\frac{\Delta(\alpha, \beta)}{\Delta(\alpha, \mu)\Delta(\beta, \mu)} \geqslant \frac{1}{2t(1-t)}.$$

But the right-hand side of this last inequality is bounded below by 2, since the function $t(1-t)$ on $[0, 1]$ takes its maximum at $t = 1/2$. So inequality (10.1.3) is established, and Theorem 10.1.1 is established in the hyperbolic case.

10.2. *Some related results*

Nathan Dunfield has shown that Theorem 10.1.1 is sharp: there exists a hyperbolic knot K in a closed orientable 3-manifold Σ with $\pi_1(\Sigma) \cong \mathbf{Z}/10\mathbf{Z}$ such that the set of boundary slopes of K has diameter 2. Furthermore, the greatest and least slopes are half-integers. (Remember that the set of slopes is defined only modulo an integer translation and a change of sign, so that the greatest and least slopes are not well-defined. On the other hand, both the diameter of the set of boundary slopes, and the condition that the greatest and least boundary slopes are integers, is invariant.) This example is presented in [22].

In this example, $\pi_1(\Sigma)$ is cyclic of even order. By contrast, Dunfield has given an argument in [25] which shows that if K is a hyperbolic knot in a closed 3-manifold Σ such that $\pi_1(\Sigma)$ is cyclic of odd order, and if the diameter of the set of all boundary slopes is exactly

2, then the greatest and least slopes cannot be integers or half-integers. The proof involves the same ingredients as Dunfield's theorem about Dehn surgery which I mentioned in Section 9.6.

Although the restrictions on the set of boundary slopes that I have stated are the only ones known for a general knot in a manifold with cyclic π_1, one can get a little more information about the set of essential surfaces in the knot exterior by looking beyond the set of slopes. For example, the main theorem of my forthcoming paper [23] with Culler gives, as a special case, information about any nontrivial knot K in a homotopy 3-sphere (e.g., S^3). Suppose that the exterior of K contains only two incompressible surfaces (so that both are bounded, one is a spanning surface and one has boundary slope $\neq 0$). If the genus g of K is $\geqslant 2$ and if s denotes the nonzero boundary slope, then

$$\frac{g}{\log_2 g} \leqslant 24s^2.$$

It's interesting to compare this with the known examples of knots in S^3 with only two essential surfaces in their exteriors, which are the torus knots: for a type (p, q) torus knot, the genus is $(p-1)(q-1)/2$ and the nonzero boundary slope is pq. So for torus knots the genus is a little less than the slope, whereas the general theorem gives an upper bound for the genus which is slightly worse-than-quadratic in the slope. (The general form of the theorem of [23], which involves the notion of a "strict boundary slope", applies to certain hyperbolic knots as well, such as the figure-eight knot.) The method of proof involves considering the functions $I_\gamma - 2$ associated to *nonperipheral* elements γ, and comparing the orders of their poles with the orders of their zeros by using some of the facts pointed out in Section 5.

In addition to the connection with classical knot theory, these results are potentially related to the Poincaré Conjecture. One can think of them as characterizations of the trivial knot in a closed orientable 3-manifold Σ with cyclic π_1: for example, Theorem 10.1.1 says that a knot in Σ is trivial if and only if the diameter of the set of its boundary slopes is strictly less than 2. If one can characterize the trivial knot, one can try to show that an *arbitrary* closed orientable 3-manifold contains some knot satisfying the condition and having an irreducible exterior. It would then follow that any closed orientable 3-manifold with cyclic π_1 would contain a trivial knot with irreducible exterior, and would therefore be a lens space. (This would give the Poincaré Conjecture as a special case.)

11. R-trees, degenerations of hyperbolic structures, and other stories

In [67], Thurston gave a criterion for the set $\mathrm{AH}(M)$ of "homotopy hyperbolic structures" on a compact irreducible 3-manifold M (with boundary) to be compact. The simplest way to define $\mathrm{AH}(M)$ involves using the $\mathrm{PSL}(2, \mathbf{C})$-character variety $PX(\pi_1(M))$, which I haven't defined in this chapter: it's the subset of $PX(\pi_1(M))$ consisting of all characters of faithful representations whose images are discrete subgroups of $\mathrm{PSL}(2, \mathbf{C})$. The question of compactness is unaffected by replacing $\mathrm{PSL}(2, \mathbf{C})$ by $\mathrm{SL}(2, \mathbf{C})$: saying that $\mathrm{AH}(M)$ is compact is equivalent to saying that the set $D(\pi_1(M)) \subset X(\pi_1(M))$ consisting of all characters of discrete faithful representations in SL(2, **C**) is compact. Thurston's theorem, which played a key role in his original proof of his geometrization theorem for Haken manifolds, asserts that this set is compact if and only if M contains no essential disks or annuli. The "only if" part is easy, and in proving the converse it's easy to see that one can assume that M contains no connected essential surfaces with nonnegative Euler characteristic.

When Marc Culler and I were doing the work that led to [20], we noticed a connection between our methods and the statement of Thurston's result. In fact, we noticed a simple proof of the weaker statement that when M contains no essential surfaces with nonnegative Euler characteristic, the intersection of $D(\pi_1(M))$ with any curve $C \subset X(\pi_1(M))$ is compact. If the conclusion were false there would be a sequence (χ_i) of points of $C \cap D(\pi_1(M))$ approaching an ideal point x of C. The discreteness and faithfulness of the representations ρ_i defining the χ_i implies that under the action on a tree T associated to x by the construction of Section 5, the stabilizer Γ_e of any edge e of T is a "small" subgroup of $\pi_1(M)$ in the sense that it contains no nonabelian free subgroup. Briefly, this is because, if γ and δ generated a free subgroup of $[\Gamma_e, \Gamma_e]$, the functions I_γ and $I_{[\gamma,\delta]}$ would take the value 2 at x by Property 5.5.3. So trace $\rho(\gamma)$, and trace $\rho([\gamma, \delta])$ would be close to 2 for large i, and this would contradict Jørgensen's inequality [39] about discrete subgroups of SL(2, **C**).

Now let's associate an essential surface F with the action of $\pi_1(M)$ on T. Since the edge stabilizers are small subgroups of $\pi_1(M)$, it follows from 2.3.1(ii) that each component of F has a small fundamental group, and hence has Euler characteristic $\geqslant 0$. This contradicts the hypothesis.

John Morgan and I were able to turn this into a proof of Thurston's compactness theorem. The task was to replace the curve C by a whole irreducible component of $X(\pi_1(M))$, and this required generalizing all the material in Sections 2, 3, and 5. In this theory, the discrete, rank-1 valuations that appear in Section 3 are replaced by more general valuations in which the "value" group **Z** is replaced by a general ordered abelian group; the simplicial trees that appear in Sections 2 and 3, are replaced by **R**-trees, which can be thought of as metric spaces in which any two points are joined by a unique topological arc; and the essential surfaces that appear in Section 2 are replaced by essential measured laminations, which can be thought of as "irrational" counterparts of essential surfaces, and which typically look locally like the product of an open set in $\mathbf{R}^2$ with a Cantor set.

Formally the two main steps in this argument were to show that a suitable kind of sequence tending to infinity in $D(\pi_1(M))$ defines an action of $\pi_1(M)$ on an **R**-tree with small arc stabilizers, and that if a 3-manifold group admits such an action then it contains an essential connected surface of nonnegative Euler characteristic. In [43], Morgan generalized the first step to hyperbolic manifolds of arbitrary dimension. Morgan and I then formulated a general conjecture about groups that act on **R**-trees with small arc stabilizers; in view of the result of [43], this conjecture implied a high-dimensional analogue of Thurston's compactness theorem. As a by-product of our proofs, we also showed that if a finitely generated 3-manifold acts *freely* on an **R**-tree then it's a free product of free groups and surface groups. We conjectured that *any* group which acts freely on an **R**-tree is of this form.

This conjecture on free actions was proved by Rips. Using Rips's ideas, Bestvina and Feighn [5] and Rips–Sela independently proved a version of our conjecture on actions with small arc stabilizers which is strong enough to imply the high-dimensional version of

Thurston's compactness theorem. In another direction, Paulin [54,55] discovered a partial generalization of the results of [45] and [43] which permit applications to much more general kinds of objects than hyperbolic manifolds. This has given rise to an entire new area of geometric group theory in which methods involving actions on **R**-trees are applied to the study of outer automorphisms, decompositions of groups, and other questions. You can learn more about this – including aspects that I haven't even mentioned, such as the connection with Thurston's compactification of Teichmüller space – from my old survey articles [60] and [61], or from Bestvina's chapter [4] in this volume.

References

[1] R.C. Alperin and H. Bass, *Length functions of group actions on Λ-trees*, Combinatorial Group Theory and Topology (Alta, Utah, 1984), Ann. of Math. Studies, Vol. 111, Princeton Univ. Press (1987), 265–378.

[2] R.C. Alperin and P.B. Shalen, *Linear groups of finite cohomological dimension*, Invent. Math. **66** (1982), 89–98.

[3] H. Bass, *Finitely generated subgroups of* GL_2, The Smith Conjecture (Papers presented at the Symposium held at Columbia University, New York, 1979), Pure and Applied Mathematics, Vol. 112, Academic Press (1984), 127–136.

[4] M. Bestvina, *ℝ-trees in topology, geometry, and group theory*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 55–91

[5] M. Bestvina and M. Feighn, *Stable actions of groups on real trees*, Invent. Math. **121** (1995), 287–321.

[6] F. Bonahon, *Geometric structures on 3-manifolds*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 219–259.

[7] S. Boyer, *Dehn surgery on knots*, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 165–218.

[8] S. Boyer and X. Zhang, *Finite Dehn surgery on knots*, J. Amer. Math. Soc. **9** (1996), 1005–1050.

[9] S. Boyer and X. Zhang, *On Culler–Shalen seminorms and Dehn filling*, Preprint.

[10] S. Boyer and X. Zhang, *A proof of the finite filling conjecture*, Preprint.

[11] E.M. Brown and R.H. Crowell, *Deformation retractions of 3-manifolds into their boundaries*, Ann. of Math. **82** (1965), 445–458.

[12] M. Brown, *Locally flat imbeddings of topological manifolds*, Ann. of Math. (2) **75** (1962), 331–341.

[13] F. Bruhat and J. Tits, *Groupes réductifs sur un corps local*, Inst. Hautes Etudes Sci. Publ. Math. **41** (1972), 5–251.

[14] D. Cooper, M. Culler, H. Gillet, D. Long, and P.B. Shalen, *Plane curves associated to character varieties of 3-manifolds*, Invent. Math. **118** (1994), 47–84.

[15] D. Cooper and D. Long, *An undetected slope in a knot manifold*, Topology '90 (Columbus, OH, 1990), Ohio State Res. Inst. Math. Publ. 1, de Gruyter (1992), 111–121.

[16] D. Cooper and D. Long, *The A-polynomial has ones in the corners*, Bull. London Math. Soc. **29** (1997), 231–238.

[17] D. Cooper and D. Long, *Representation theory and the A-polynomial of a knot*, Knot Theory Appl. Chaos Solitons Fractals **9** (1998), 749–763.

[18] M. Culler, C. Gordon, J. Luecke and P.B. Shalen, *Dehn surgery on knots*, Ann. of Math. **125** (1987), 237–300.

[19] M. Culler and J.W. Morgan, *Group actions on* **R**-*trees*, Proc. London Math. Soc. (3) **55** (1987), 571–604.

[20] M. Culler and P.B. Shalen, *Varieties of group representations and splittings of 3-manifolds*, Ann. of Math. **117** (1983), 109–146.

[21] M. Culler and P.B. Shalen, *Bounded separating, incompressible surfaces in knot manifolds*, Invent. Math. **75** (1984), 537–545.

[22] M. Culler and P.B. Shalen, *Boundary slopes of knots*, Comment. Math. Helv., to appear.

[23] M. Culler and P.B. Shalen, *On knots with only two essential surfaces*, In preparation.

[24] M. Davis and J.W. Morgan, *Finite group actions on homotopy 3-spheres*, The Smith Conjecture (Papers presented at the Symposium held at Columbia University, New York, 1979), Pure and Applied Mathematics, Vol. 112, Academic Press (1984), 181–225.

[25] N. Dunfield, *Examples of non-trivial roots of unity at ideal points of hyperbolic 3-manifolds*, Topology, to appear.

[26] N. Dunfield, *Cyclic surgery, degrees of maps of character curves, and volume rigidity of hyperbolic manifolds*, Preprint.

[27] M.J. Dunwoody, *An equivariant sphere theorem*, Bull. London Math. Soc. **17** (1985), 437–448.

[28] A.L. Edmonds, *A topological proof of the equivariant Dehn lemma*, Trans. Amer. Math. Soc. **297** (1986), 605–615.

[29] D.B.A. Epstein, *Curves on 2-manifolds and isotopies*, Acta Math. **115** (1966), 83–107.

[30] C.D. Feustel, *A splitting theorem for closed orientable 3-manifolds*, Topology **11** (1972), 151–158.

[31] W. Fulton, *Algebraic Curves. An Introduction to Algebraic Geometry*, Advanced Book Classics, Addison-Wesley (1989).

[32] F. González-Acuña and J.M. Montesinos-Amilibia, *On the character variety of group representations in* $SL(2, \mathbf{C})$ *and* $PSL(2, \mathbf{C})$, Math. Z. **214** (1993), 627–652.

[33] A.E. Hatcher, *On the boundary curves of incompressible surfaces*, Pacific J. Math. **99** (1982), 373–377.

[34] J. Hempel, *3-Manifolds*, Ann. of Math. Studies, Vol. 86, Princeton Univ. Press (1976)

[35] J.F.P. Hudson, *Piecewise Linear Topology*, University of Chicago Lecture Notes prepared with the assistance of J.L. Shaneson and J. Lees, W.A. Benjamin, New York (1969).

[36] W. Jaco and J.H. Rubinstein, *PL equivariant surgery and invariant decompositions of 3-manifolds*, Adv. in Math. **73** (1989), 149–191.

[37] W. Jaco and P.B. Shalen, *Seifert fibered spaces in 3-manifolds*, Mem. Amer. Math. Soc. **21** (220) (1979).

[38] K. Johannson, *Homotopy Equivalences of 3-Manifolds with Boundaries*, Lecture Notes in Mathematics, Vol. 761, Springer, Berlin (1979).

[39] T. Jørgensen, *On discrete groups of Möbius transformations*, Amer. J. Math. **98** (1976), 739–749.

[40] A.G. Kurosh, *The Theory of Groups*, 2nd edn, Chelsea (1960).

[41] W.H. Meeks III and S.T. Yau, *The equivariant loop theorem for three-dimensional manifolds and a review of the existence theorems for minimal surfaces*, The Smith Conjecture (Papers presented at the Symposium held at Columbia University, New York, 1979), Pure and Applied Mathematics, Vol. 112, Academic Press (1984), 153–163.

[42] E.E. Moise, *Geometric Topology in Dimensions 2 and 3*, Graduate Texts in Mathematics, Vol. 47, Springer, Berlin (1977).

[43] J.W. Morgan, *Groups actions on trees and the compactification of the space of classes of* $SO(n, 1)$*-representations*, Topology **25** (1986), 1–33.

[44] J. Morgan and H. Bass, eds, *The Smith Conjecture*, Papers presented at the Symposium held at Columbia University, New York, 1979, Pure and Applied Mathematics, Vol. 112, Academic Press (1984).

[45] J. Morgan and P.B. Shalen, *Degenerations of hyperbolic structures, I*, Ann. of Math. **120** (1984), 401–476.

[46] J. Morgan and P.B. Shalen, *Degenerations of hyperbolic structures, II*, Ann. of Math. **127** (1988), 403–456.

[47] J. Morgan and P.B. Shalen, *Degenerations of hyperbolic structures, III*, Ann. of Math. **127** (1988), 457–519.

[48] D. Mumford, *Algebraic Geometry. I. Complex Projective Varieties*, Grundlehren der Mathematischen Wissenschaften, Vol. 221, Springer, Berlin (1976).

[49] L.P. Neuwirth, *Interpolating manifolds for knots in* S^3, Topology **2** (1963), 359–365.

[50] L.P. Neuwirth, *Knot Groups*, Ann. of Math. Studies, Vol. 56, Princeton Univ. Press (1965).

[51] I. Niven, H.S. Zuckerman and H.L. Montgomery, *An Introduction to the Theory of Numbers*, 5th edn, Wiley (1991).

[52] J.-P. Otal, *Le Théorème d'Hyperbolisation pour les Variétés Fibrées de Dimension 3*, Astérisque, Vol. 235, Société Mathématique de France (1996).

[53] J.-P. Otal, *Thurston hyperbolization of Haken manifolds*, Surveys in Differential Geometry, S.-T. Yau ed., (1997), to appear.

[54] F. Paulin, *Topologie de Gromov équivariante, structures hyperboliques et arbres réels*, Invent. Math. **94** (1988), 53–80.

[55] F. Paulin, *Outer automorphisms of hyperbolic groups and small actions on* $\mathbf{R}$*-trees*, Arboreal Group Theory (Berkeley, CA, 1988), MSRI Publ. 19, Springer, New York (1991), 331–343.

[56] P. Scott and C.T.C. Wall, *Topological methods in group theory*, Homological Group Theory (Proc. Sympos. Durham 1977), London Math. Soc. Lecture Note Series, Vol. 37, Cambridge Univ. Press (1979), 137–203.
[57] J.-P. Serre, *A Course in Arithmetic*, Graduate Texts in Mathematics, Vol. 7, Springer, Berlin (1973) (Translated from the French).
[58] J.-P. Serre, *Trees*, Springer, Berlin (1980) (Translated from the French by J. Stillwell).
[59] I.R. Shafarevich, *Basic Algebraic Geometry*, Springer, Berlin (1977) (Translated from the Russian by K.A. Hirsch).
[60] P.B. Shalen, *Dendrology of groups*, Essays in Group Theory, S. Gersten, ed., M.S.R.I. Pub. 8, Springer, Berlin (1987), 265–320.
[61] P. Shalen, *Dendrology and its applications*, Group Theory from a Geometrical Viewpoint, ICTP, Trieste, Italy, 26 March–6 April, 1990, E. Ghys, A. Haefliger and A. Verjovsky, eds, World Scientific Publishing (1991), 543–616.
[62] E.H. Spanier, *Algebraic Topology*, Springer, Berlin (1981).
[63] J.R. Stallings, *On fibering certain 3-manifolds*, Topology of 3-Manifolds and Related Topics (Proc. Univ. of Georgia Institute 1961), Prentice-Hall (1962), 95–100.
[64] J.R. Stallings, *A topological proof of Grushko's theorem on free products*, Math. Z. **90** (1965), 1–8.
[65] J. Stallings, *Group theory and three-dimensional manifolds*, A James K. Whittemore Lecture in Mathematics given at Yale University, 1969, Yale Mathematical Monographs, Vol. 4, Yale University Press (1971).
[66] W.P. Thurston, *A norm for the homology of 3-manifolds*, Mem. Amer. Math. Soc. **59** (339) (1986), 1–viii and 99–130.
[67] W.P. Thurston, *Hyperbolic structures on 3-manifolds. I. Deformation of acylindrical manifolds*, Ann. of Math. (2) **124** (1986), 203–246.
[68] A. Weil, *On discrete subgroups of Lie groups*, Ann. of Math. (2) **72** (1960), 369–384.
[69] A. Weil, *On discrete subgroups of Lie groups, II*, Ann. of Math. (2) **75** (1962), 578–602.
[70] F. Waldhausen, *Gruppen mit Zentrum und 3-dimensionale Mannigfältigkeiten*, Topology **6** (1967), 505–517.

CHAPTER 20

Topological Rigidity Theorems

Christopher W. Stark*

Division of Mathematical Sciences, National Science Foundation, 4201 Wilson Boulevard, Room 1025, Arlington, VA 22230, USA

Contents

---

*Partially supported by an NSA grant.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

1. Introduction

Two topological manifolds which are properly homotopy equivalent may easily fail to be homeomorphic. Since the 1930s the lens spaces $L(7,1)$ and $L(7,2)$ have been known to be homotopy equivalent but not simple homotopy equivalent [29], for example, and the topological invariance of Whitehead torsion [24] shows that manifolds which are not simple homotopy equivalent are not homeomorphic.

If we can prove for a manifold M^n that any manifold which is properly homotopy equivalent to M is homeomorphic to M then we say that M^n is topologically rigid, by analogy to results in geometry which assert that any deformation of a geometric structure is equivalent to the original. Obstructions to modifying a homotopy equivalence to a homeomorphism are found in the Whitehead group and in surgery theory, and rigidity arguments tend to mingle geometry and algebra to show that these obstructions vanish. Results in the subject usually specialize to a geometrically defined class of manifolds.

Most topological rigidity results concern manifolds which are aspherical (connected, with contractible universal covering spaces). The homotopy theory of these spaces is particularly simple: a continuous map between aspherical CW complexes is a homotopy equivalence if and only if it induces an isomorphism of fundamental groups. The primal example of an aspherical manifold is Euclidean space; the most familiar closed aspherical manifolds are the tori $T^n = (S^1)^n$.

CONJECTURE 1.1 (*Borel Conjecture*). Every homotopy equivalence $f : M \to N$ of closed aspherical manifolds is homotopic to a homeomorphism.

Conjecture 1.1 is reported by F.T. Farrell and W.-C. Hsiang [108] to have entered the oral culture by the 1960s as A. Borel's response to [134]. In that study of solvmanifolds (coset spaces $\Gamma \backslash S$, where S is a solvable Lie group) Mostow showed that if two closed solvmanifolds have isomorphic fundamental groups then they are diffeomorphic. By the early to middle 1960s it was known that the connected sum of an exotic sphere and a torus may carry a nonstandard differentiable structure [17, Corollary 2.8], so smooth rigidity could not be expected. The classification of PL homotopy tori in the late 1960s (see Section 2.3) rules out piecewise linear rigidity claims.

The truth of Conjecture 1.1 for surfaces is classical (see Section 2.1) and by the late 1960s it was known that if two irreducible, sufficiently large 3-manifolds are homotopy equivalent then they are homeomorphic [169]. Surgery theory is the technology relevant to the rigidity problems in high dimensions, with satellite algebraic problems on K- and L-groups, and the following generalization of Conjecture 1.1 arises naturally when the rigidity conjecture is translated into a claim about the long exact sequence of topological surgery.

CONJECTURE 1.2 (*Generalized Borel Conjecture*). If $f : (M, \partial M) \to (N, \partial N)$ is a homotopy equivalence of compact aspherical manifolds which restricts to a homeomorphism of boundaries then f is homotopic rel boundary to a homeomorphism.

This generalization is more and more frequently referred to as "the Borel Conjecture", because of parallels with the Novikov Conjecture and because surgical methods usually yield Conjecture 1.2 if they prove Conjecture 1.1. Some work has also been done on versions of these problems for open manifolds: if noncompact aspherical manifolds are properly homotopy equivalent, are they homeomorphic?

Another problem concerning the topology of aspherical manifolds was discussed in 1969 by Wall [176] as a mate to the spherical spaceform problem [166,127]:

PROBLEM 1.1 (*The topological Euclidean space form problem*). Classify all free proper actions of discrete groups on Euclidean space.

This is a descendant of Hilbert's Eighteenth Problem on crystallographic groups [133], which was solved by Bieberbach (see Section 4.1). Problem 1.1 is usually understood to concern both existence and uniqueness. A rigidity or uniqueness claim in this context would conclude that two actions of a group Γ on $\mathbb{R}^n$ are topologically conjugate and would be a case of the Borel Conjecture or its proper counterpart.

Existential results should assume algebraic properties of the group Γ and deduce that there is a free proper action of Γ on some $\mathbb{R}^n$; they might also give information about appropriate values of n and the action (we especially want to know if $\Gamma\backslash\mathbb{R}^n$ may be a closed manifold). See Sections 3.1 and 9 for some results and a statement of the existence problem.

After M. Davis produced examples of closed manifolds whose universal covering spaces are contractible but non-Euclidean [38] the scope of investigations into aspherical manifolds enlarged, and developments in geometric group theory and on the Novikov conjecture have further altered our view of the rigidity problem.

2. Examples and classical facts

Positive results in a few cases of low dimension or tractable fundamental group shaped later approaches to rigidity problems. This section sketches the cases of 2-manifolds and n-dimensional tori, as well as the most important simply connected manifolds. Work on 3-manifolds has also influenced high-dimensional arguments, especially splitting theorems, and those results are reported in later sections.

2.1. *Example 1: Surfaces*

The classification of closed surfaces is often sketched in elementary topology courses and yields the strongest rigidity conclusions known in any dimension, since asphericity is not assumed: If closed 2-manifolds M and N are homotopy equivalent then they are homeomorphic. Moreover, any homotopy equivalence of closed surfaces deforms to a homeomorphism.

We may argue for the statements above with an outline going back to Riemann: dissect a target surface along nicely imbedded circles, use transversality to dissect the source manifold along preimages of these circles, and then modify the map to simplify it towards a homeomorphism. This line of attack, sometimes called splitting, is a mainstay of

3-manifold topology [169] and has also been used and studied in higher dimensions, where it may be obstructed (see Section 6).

Riemann surfaces with elliptic points are orbifolds rather than manifolds, but have rigidity properties similar to those of 2-manifolds when we consider maps which preserve singular data (stratified or equivariant maps). Once again, complex analysis and geometry provide better information on stratified rigidity questions for 2-orbifolds than we expect to find in other dimensions (Section 10.2).

2.2. *Classical results*

Several classical facts guide and constrain rigidity theorems in higher dimensions. Rigid manifolds with nontrivial fundamental group might be expected to have rigid universal covering spaces (although we are learning that less control over the rigidifying homeomorphisms might be possible than one would like [5]), so we consider here the simplest compact and noncompact manifolds.

Milnor's discovery of exotic smooth structures on spheres [132] led to the classification of manifold structures of different kinds. Topological versions of tools familiar from Smale's generalized Poincaré conjecture show that there are no exotic topological structures on spheres, except perhaps in dimension three (see [154, Section 4.13] for $n \geqslant 5$ and [92, Chapter 7] for $n = 4$):

THEOREM 2.1 (Topological rigidity of spheres). *Let $n \neq 3$. If a closed manifold M^n is homotopy equivalent to S^n then M^n is homeomorphic to S^n.*

Note that degree theory shows that every homotopy equivalence on an n-sphere is homotopic to a homeomorphism, so spheres are topologically rigid except perhaps in dimension 3.

However, there are homotopy types which contain many non-homeomorphic closed and simply connected manifolds, and these are classified by surgery theory (see Section 5.2 and the chapter of this volume on surgery). In addition, closed and simply connected manifolds may admit self-homotopy equivalences which do not deform to homeomorphisms (a sample computation with modern tools, following an idea which goes back to Novikov, appears in [149, Example 20.4]).

Stallings's characterization of Euclidean space [163] leads to a counterpart to the topological rigidity of spheres.

THEOREM 2.2 (Characterization of Euclidean space). *Let $n \geqslant 5$. If M^n is a noncompact contractible manifold which is simply connected at infinity then M^n is homeomorphic to $\mathbb{R}^n$.*

The characterization theorem, the invariance of simple connectedness at infinity under proper homotopy equivalence, and proper degree theory yield a rigidity statement.

COROLLARY 2.1 (Rigidity of Euclidean space). *If $f : M^n \to \mathbb{R}^n$ is a proper homotopy equivalence of open manifolds and $n \geqslant 5$ then f is properly homotopic to a homeomorphism.*

2.3. *Example 2: The torus*

One of the simplest noncontractible aspherical manifolds, the n-dimensional torus, illustrates the range of issues which appear under the heading of rigidity. We begin with the properties of the torus T^n as a flat Riemannian manifold.

Let g_0 denote the Riemannian metric on a product of n circles of radius 1. If g is any other flat Riemannian metric on T^n then there is an affine diffeomorphism of T^n which transforms g_0 to g, so up to affine equivalence, there is only one flat n-torus (see Section 4). By considering the lengths of shortest closed geodesics one sees easily that there are uncountably many isometrically distinct flat Riemannian metrics on T^n. The isometry classification of two-dimensional flat tori presented in elementary complex analysis is a model for the first steps in the study of deformations and rigidity for discrete subgroups of Lie groups (also Section 4).

The classification of PL structures on T^n for $n \geqslant 5$ was an early success of surgery on manifolds with nontrivial fundamental group [109,174,175]. This result of Wall and Hsiang–Shaneson played an essential role in the last stages of the Kirby–Siebenmann theorem on triangulating topological manifolds [121]. Although the set of PL structures on T^n contains more than one element, passing to a 2^n-sheeted covering space renders all these PL structures equivalent. Smooth structures on the torus have a similar lifting property, and it seems to be an open question whether distinct PL (respectively smooth) structures on most aspherical manifolds should agree after lifting to some finite-sheeted covering space.

Any topological manifold homotopy equivalent to T^n must be homeomorphic to the standard torus. The extension of surgery theory from the smooth and PL categories to the category of topological manifolds (see Section 5) leads to an easy proof of this assertion using splitting theorems of Bass–Heller–Swan and Shaneson (Section 6). In fact, more is known [122, pp. 264–283]:

THEOREM 2.2 (Rigidity of tori). *Let $f : M^{n+k} \to T^n \times D^k$ be a homotopy equivalence which restricts to a homeomorphism of the boundaries. If $n + k \geqslant 5$ then f is homotopic to a homeomorphism.*

A quotient space of T^n by a finite group acting smoothly and freely admits a flat Riemannian metric, and all flat Riemannian manifolds admit such a description. The analysis of flat manifolds by Farrell and Hsiang (Section 7.3) mixes Frobenius reciprocity (Section 5.4) and a circle of ideas known as controlled topology. The covering projection $T^n \to T^n$ which has degree $r > 0$ in each circle factor has counterparts for every closed Riemannian flat manifold. These expanding self maps suggest vanishing proofs which represent algebraic data by geometric constructions with associated size information, assuming that the algebraic element under consideration is trivial if a size estimate is sufficiently small: this is the idea at the core of controlled topology (Section 7).

If a finite group Q acts smoothly and properly discontinuously but not freely on T^n then the quotient space is a flat orbifold. Structure and rigidity results in this situation may consider T^n/Q as a stratified space or may emphasize the equivariant topology of the action (Q, T^n). We have quite a bit of information about the flat case (see especially [35,36]) but still have a lot to do in general (Section 10).

3. Constructions and basic properties of aspherical manifolds

3.1. *Basic properties*

The fundamental groups of aspherical manifolds have a number of algebraic properties. If M^n is any aspherical manifold then its fundamental group has cohomological dimension n or less. This is the only constraint in the most general case:

THEOREM 3.1 (Johnson [176]). *Let Γ be a countable discrete group. Γ is of finite cohomological dimension if and only if there is a proper smooth free action of Γ on $\mathbb{R}^n$ for all sufficiently large n.*

If M^n is compact then $\Gamma = \pi_1(M)$ satisfies additional constraints. Because a compact manifold has the homotopy type of a finite CW complex [122], the fundamental group Γ of such a manifold is of type FL [18,156].

DEFINITION 3.1. Let Γ be a countable group.

(a) Γ is of *type FL* if and only if the trivial $\mathbb{Z}\Gamma$ module $\mathbb{Z}$ has a finite-length resolution by finitely generated free $\mathbb{Z}\Gamma$-modules.

(b) Γ is of *type FP* if and only if the trivial $\mathbb{Z}\Gamma$ module $\mathbb{Z}$ has a finite-length resolution by finitely generated projective $\mathbb{Z}\Gamma$-modules.

If Γ is the fundamental group of a closed aspherical manifold then, in addition to being of type FL, Γ is a Poincaré duality group: we will return to this condition in Section 9. The fundamental groups of some noncompact aspherical manifolds lie in a larger class, known as duality groups; nonuniform arithmetic subgroups of semisimple linear Lie groups are the core examples. K. Brown's textbook [18] is a good guide to the finiteness properties cited above, along with duality groups and Poincaré duality groups.

3.2. *Coset spaces and curvature conditions*

Differential geometry is the most classical source of aspherical manifolds, mainly through the following result:

THEOREM 3.2 (The Cartan–Hadamard theorem). *If V^n is a simply connected smooth manifold with a complete Riemannian metric of nonpositive sectional curvature then for every $v \in V$ the exponential map* $\exp\colon T_vV \to V$ *is a diffeomorphism.*

COROLLARY 3.1. *Every complete Riemannian manifold of nonpositive sectional curvature is aspherical.*

The almost flat manifolds studied by Gromov [97] are also aspherical, as are many almost nonpositively curved manifolds [94]. Gromov's characterization of almost flat manifolds leads into Lie groups and coset spaces:

DEFINITION 3.2. A smooth manifold M^n is *almost flat* if it admits a sequence of complete Riemannian metrics g_i and a number $D > 0$ such that for all i we have an upper bound $\operatorname{diam}(M, g_i) \leqslant D$, while the sectional curvatures of the g_i converge to 0 uniformly.

DEFINITION 3.3. An *infranilmanifold* is a double coset space $M = \Gamma\backslash G/K$, where G is a Lie group which is virtually connected and virtually nilpotent, K is a maximal compact subgroup of G, and Γ is a discrete subgroup of G.

THEOREM 3.3 (Gromov [97]). *A smooth manifold is almost flat if and only if it is an infranilmanifold.*

More generally, if G is any virtually connected Lie group and $K < G$ is a maximal compact subgroup then G/K is diffeomorphic to a Euclidean space, while any discrete subgroup $\Gamma < G$ acts properly discontinuously by left translation on G/K. If Γ is torsion-free then this is a free action and $\Gamma\backslash G/K$ is an aspherical manifold with fundamental group Γ, while if Γ contains elements of finite order then these elements fix points of G/K and $\Gamma\backslash G/K$ is an orbifold.

The subgroup Γ of the Lie group G is said to be *uniform* or *cocompact* if $\Gamma\backslash G$ is compact in the quotient topology, while Γ is said to be a *lattice* in G if $\Gamma\backslash G$ has finite volume with respect to any left-invariant volume form on G. (Note that both these properties depend upon the imbedding of Γ in G.) Two subgroups A and B of a group G are said to be *commensurable* if $A \cap B$ is of finite index in both A and B.

DEFINITION 3.4. Let G be an algebraic subgroup of GL_n, defined over the field $\mathbb{Q}$ of rational numbers. Thus, G is defined by a set $\mathcal{S}$ of polynomial equations in the entries of $n \times n$ matrices and in the inverse of the determinant, and these equations have rational coefficients.

(a) If $\mathbb{F}$ is any extension field of $\mathbb{Q}$ then $G_{\mathbb{F}}$ denotes the set of solutions of $\mathcal{S}$ with entries in $\mathbb{F}$. ($G_{\mathbb{F}}$ is known as the set of $\mathbb{F}$-points of G; it is a subgroup of $\mathrm{GL}_n(\mathbb{F})$.)
(b) $G_{\mathbb{Z}}$ denotes $G_{\mathbb{Q}} \cap \mathrm{GL}_n(\mathbb{Z})$.
(c) A subgroup Γ of $G_{\mathbb{Q}}$ is *arithmetic* if Γ and $G_{\mathbb{Z}}$ are commensurable.
(d) A group Δ is arithmetic if it can be imbedded as an arithmetic subgroup of $G_{\mathbb{Q}}$ for some $\mathbb{Q}$-arithmetic subgroup G of GL_n.

For example, the group $\mathrm{SL}_n(\mathbb{Z})$ of matrices with integral entries and determinant 1 is an arithmetic subgroup of $\mathrm{SL}_n(\mathbb{Q})$. Serre's survey [157] is highly recommended to topologists looking into arithmetic groups, along with the relevant sections of Brown's book [18]. Note that $\mathrm{SL}_n(\mathbb{Z})$ contains elements of finite order, such as permutation matrices; this is common in arithmetic groups, but they are nonetheless a source of aspherical manifolds thanks to the following result (see [155], [15, pp. 113–115], or [147, pp. 93–95]):

THEOREM 3.4 (Selberg's lemma). *If Δ is a finitely generated subgroup of* $\mathrm{GL}_n(\mathbb{C})$ *then Δ contains a torsion-free subgroup of finite index.*

The discrete subgroups constructed in the proof of the next theorem [15] are arithmetic:

THEOREM 3.5 (Borel). *Every connected semisimple Lie group G contains a discrete uniform subgroup.*

The work of Margulis shows that in many semisimple Lie groups every lattice is arithmetic with respect to some $\mathbb{Q}$-structure [129,130]. Much earlier, Mal'cev had done as much for lattices in nilpotent Lie groups [128].

3.3. *Torus actions and singular fiberings*

If K is a compact, connected Lie group acting effectively on an aspherical manifold then K must be a torus (see Theorem 10.1). Constructions of torus actions are also an effective means of producing aspherical manifolds, often as twisted products

$$M^{n+k} = V^k \times_Q T^n,$$

where V is an aspherical manifold and Q is a discrete group.

These actions are closely related to the Seifert fibered spaces of 3-manifold topology and have been thoroughly studied by Conner, Raymond, Lee, and others [31,32,123,124]. Singular fibering constructions have also been studied with fibers other than tori [117,151].

The features of singular fiberings which have been relevant for rigidity investigations are associated to combinatorial descriptions of these spaces as a stratified system of bundles over a base orbifold of lower dimension, and to the topology of the base orbifold. Constructability questions often pass through an obstruction in group cohomology, which Conner, Lee, and Raymond have analyzed to good effect. See the chapter in this volume by Lee and Raymond for more detailed information.

3.4. *Groups generated by reflections*

Section 1 cited the work of M. Davis on aspherical manifolds constructed from groups generated by reflections. We sketch this important construction here and recommend the chapter by Davis in this volume as well as the papers [38,39].

The basic version of these constructions begins with a manifold with triangulated boundary, $(X, \partial X)$ and associates to the top-dimensional simplices of ∂X generators for a Coxeter system; copies of X are glued together along these simplices to produce a space U on which Γ acts with fundamental domain X. Coxeter groups had long been considered in spaces of constant curvature [16,114]; see especially the work of Vinberg [168].

A *Coxeter system* is a pair (Γ, V), where Γ is a group and V is a set of elements $v_i \in \Gamma$ such that Γ has a presentation in terms of the elements v_i and $m(i, j) \in \mathbb{Z}\ (> 0) \cup \{\infty\}$

$$\Gamma = \langle v_i \in \Gamma\colon \text{ for each } i,\ v_i^2 = e \text{ and for } i \neq j\ (v_i v_j)^{m(i,j)} = e \text{ if } m(i, j) < \infty \rangle.$$

If a group Γ admits such a description then we say that it is a Coxeter group. The space U is constructed by this method: if X is a space with a family of subspaces X_i indexed by a set V and if (Γ, V) is a Coxeter system with generators V then we define a space $U = (\Gamma \times X)/{\sim}$, where the equivalence relation is generated by $(g, x) \sim (h, y)$ if and only if $x = y$, $x \in X_i$ and $g^{-1}h = \gamma_i$. Γ acts on U by left translation in the Γ factor, and $\Gamma \backslash U \cong X$.

Davis applies this construction to a contractible manifold X with boundary; the resulting space $U = U(X, \Gamma)$ is then a contractible manifold if the triangulation of ∂X is sufficiently fine, Γ is an infinite Coxeter group, and each $m(i, j)$ is finite (many of the well-studied examples take $m(i, j) = 2$). If ∂X is a nonsimply connected homology sphere then U is often not simply connected at infinity, so that U is not homeomorphic to Euclidean space, by Theorem 2.2. Because every Coxeter group Γ has a torsion-free subgroup Γ_0 of finite index, Davis produces a closed manifold $M = \Gamma_0 \backslash U$ with contractible, non-Euclidean universal covering space U. The reader should be aware that this sketch has omitted some important details in the proof of the following result [38]:

THEOREM 3.6 (Davis's aspherical manifolds). *In every dimension greater than or equal to four there exist closed aspherical manifolds not covered by Euclidean space.*

The construction of U sketched above reappears in the argument for the following useful observation [39, pp. 213–215].

THEOREM 3.7 (Davis's doubling trick). *If X is an aspherical manifold with triangulated boundary then X is a retract of a closed aspherical manifold.*

Note that by taking regular neighborhoods in Euclidean space we see that there exists a compact aspherical manifold with fundamental group Γ if and only if Γ is the fundamental group of a finite, aspherical simplicial complex. This consequence of Theorem 3.7 appears in [39, pp. 215–217]:

THEOREM 3.8. *If the Novikov Conjecture (respectively the Integral Novikov Conjecture) holds for the fundamental group of every closed aspherical manifold then it holds for every group Γ admitting a finite $K(\Gamma, 1)$.*

Bizhong Hu has used similar arguments for fundamental groups of polyhedra of nonpositive curvature in the sense of Alexandrov [112]. Moussong's work [137] on nonpositively curved polyhedral metrics for Coxeter complexes leads to contractible topological manifolds V^n with complete polyhedral metrics of nonpositive curvature such that V^n is not homeomorphic to $\mathbb{R}^n$. Note also that these examples show that the polyhedral counterpart of the Cartan–Hadamard Theorem (3.2 above) is false.

See Section 10.2 for rigidity results in this context.

3.5. *Polyhedra of nonpositive curvature*

Throughout this section we consider curvature in the sense of Alexandrov and Toponogov, which applies to polyhedra rather than manifolds, as in Gromov's influential paper on hyperbolic groups [98]. See the chapter in this volume by Davis for details.

Gromov introduced a construction called *hyperbolization*, which modifies a polyhedron to produce a polyhedron which is nonpositively curved. This idea has been worked out with care and some improvements by Davis and Januszkiewicz [41] and Charney and Davis [28]: we warn the reader that Gromov's original treatment of relative hyperbolization is not adequate for applications and Charney–Davis should be consulted for details.

Hyperbolization constructions proceed by a cell replacement construction: every k-cell in a polyhedron is replaced by a canonical space of the same dimension which has non-trivial fundamental group and the combinatorics of the polyhedron are reproduced in the attaching data for these canonical pieces. Some observations about this kind of construction are elementary: Lemma 3.1 on aspherical pasting shows that if every canonical piece is aspherical and the attaching maps are built with a little care and induce injections on fundamental groups then we should produce an aspherical polyhedron modeled on the combinatorics of the original complex. It is also not surprising that this construction can be arranged so that the n-dimensional canonical piece has a preferred map onto an n-cell, and so that the hyperbolized polyhedron has an essential map onto the original space. Considerable attention is required to manage curvature during the rebuilding process, especially in relative hyperbolization, which leaves a subcomplex intact during the enlargement process.

The first contribution of hyperbolization is statistical: a large number of groups are fundamental groups of finite polyhedra with nonpositive or negative curvature, including many triangulated manifolds or Poincaré complexes. Another application, which depends on relative hyperbolization, was announced by Gromov and improved by Charney and Davis [28]:

THEOREM 3.9. *Every triangulable manifold is cobordant to a triangulable manifold of strictly negative curvature.*

3.6. *Aspherical complements*

A number of open manifolds obtained by deletion operations are known to be aspherical. We include pointers to some of these spaces, since any proper version of the Borel Conjecture must encompass them.

Let K be a polyhedral circle in S^3. The asphericity of the knot complement $S^3 \setminus K$ was the testbed and motivation for some of the fundamental results of 3-manifold topology [106,152].

Following work of Arnol'd and Deligne, many hyperplane arrangements in Euclidean space have been shown to have aspherical complements [140, Chapter 5].

J.H.C. Whitehead exhibited a contractible open manifold which is not Euclidean space by taking the complement of a certain compactum in the 3-sphere [152,184].

3.7. *Moduli spaces of surfaces*

A class of examples arises in Riemann surface theory which is closely related to manifolds of nonpositive curvature, but which seems not to overlap with them.

Let Σ_g^2 be a closed, orientable surface of genus g and let $\mathcal{T}_g$ denote the Teichmüller space of marked hyperbolic structures on Σ_g^2; an important result in surface theory asserts that $\mathcal{T}_g \cong \mathbb{R}^{6g-g}$. The group of outer automorphisms of the fundamental group of our surface is the quotient of the group of all automorphisms by the normal subgroup of inner automorphisms, $\mathrm{Out}(\pi_1(\Sigma_g^2)) = \mathrm{Aut}(\pi_1(\Sigma_g^2))/\mathrm{Inn}(\pi_1(\Sigma_g^2))$. There is an action of $\mathrm{Out}(\pi_1(\Sigma_g^2))$ on $\mathcal{T}_g$ which changes markings; this action is properly discontinuous, but not free, since the group of outer automorphisms contains finite subgroups.

$\mathrm{Out}(\pi_1(\Sigma_g^2))$ has many of the properties of arithmetic groups. In particular, this group has a torsion-free subgroup Γ of finite index, and $\mathcal{T}_g/\Gamma$ is an open aspherical manifold which is known to have good compactifications [101–103].

The mapping class group $\mathrm{Out}(\pi_1(\Sigma_g^2))$ is studied in algebraic geometry and geometric group theory, as well as in the topology of aspherical manifolds.

More generally, if M^n is any aspherical manifold then $\mathrm{Out}(\pi_1(M^n))$ acts on the structure set $\mathcal{S}^{\mathrm{TOP}}(M^n)$ (see Definition 5.1) and plays an important role in the study of the group of homeomorphisms of M^n [32].

3.8. *Branched coverings and pasting constructions*

Constructions which alter or combine aspherical manifolds to produce new ones include the formation of fiber bundles with aspherical base and fiber, certain group actions (see Section 3.3), and the two constructions discussed below.

Branched coverings are familiar in the geometry of surfaces and 3-manifolds. They were used by Gromov and Thurston [100] in higher dimensions to build manifolds with metrics of nonpositive curvature which are close to hyperbolic manifolds although they admit no Riemannian metric of constant negative sectional curvature.

Pasting constructions can be done within the class of aspherical complexes, using the following lemma which goes back to [184].

LEMMA 3.1 (Aspherical pasting lemma). *Suppose that X is a finite complex such that each component of X is aspherical. Suppose that A_0 and A_1 are aspherical subcomplexes of X such that each inclusion $h_i : A_i \hookrightarrow X$ induces an injection on the fundamental group. If there is a homotopy equivalence $\phi : A_0 \to A_1$ then every component of the adjunction space $Y = X \cup (A_0 \times [0, 1])$ formed by attaching $A_0 \times \{0\}$ to X by h_0 and attaching $A_0 \times \{1\}$ to X by $h_1 \circ \phi$ is aspherical.*

This result suggests the splitting strategy explored in Section 6 and gives the fundamental groupoid of Y the structure of an HNN extension or free product with amalgamations (i.e., the fundamental groupoid of a graph of groups). For this reason, work with pasting and splitting of aspherical manifolds tends to have close connections with combinatorial group theory.

Gromov and Piatetski-Shapiro used a pasting construction to hybridize two arithmetic groups and produce new non-arithmetic hyperbolic manifolds [99]. (See also the discussion in [130, Appendix C].)

4. Rigidity theorems in geometry

Two celebrated theorems provide models for a number of rigidity results in geometry and equivariant topology. Bieberbach's solution to Hilbert's Eighteenth Problem on crystallographic groups begins with metric data, namely a flat Riemannian metric, and asserts that a homotopy equivalence determines metric structure up to a weaker equivalence relation, namely affine isomorphism. (A flat metric can be rescaled by a constant, so affine equivalence is the tightest geometric equivalence relation which might coincide with homotopy equivalence.) Mostow's rigidity theorem for hyperbolic manifolds deforms a homotopy equivalence to an isometry, but the most natural setting for the theorem works with discrete subgroups of Lie groups rather than metric structures on manifolds.

4.1. *Bieberbach's theorem*

THEOREM 4.1 (Bieberbach). *If $f: M \to N$ is a homotopy equivalence between closed, connected, flat Riemannian manifolds then f is homotopic to an affine diffeomorphism.*

A diffeomorphism of flat Riemannian manifolds is affine if and only if it carries geodesics to geodesics; we amplify to give a global view of Bieberbach's theorem. The universal covering space $\widetilde{M}^n$ of a complete flat Riemannian manifold M^n is isometric to Euclidean space in the standard metric, and the action of the fundamental group of M^n on $\widetilde{M}^n$ by deck transformations is identified with the action of a discrete subgroup of the group $\mathrm{E}(n)$ of rigid Euclidean motions on $\mathbb{R}^n$. Theorem 4.1 is part of the structure theory of crystallographic groups [185, Chapter 3] developed by Bieberbach, which builds upon the fact that $\mathrm{E}(n)$ is a split extension

$$1 \longrightarrow \mathbb{R}^n \longrightarrow \mathrm{E}(n) \longrightarrow \mathrm{O}(n) \longrightarrow 1, \tag{4.1}$$

in which $\mathrm{O}(n)$ acts on $\mathbb{R}^n$ by the tautological representation. The corresponding group $\mathrm{Aff}(n)$ of affine motions enlarges the quotient in the split extension:

$$1 \longrightarrow \mathbb{R}^n \longrightarrow \mathrm{Aff}(n) \longrightarrow \mathrm{GL}(n) \longrightarrow 1,$$

and the inclusion $\mathrm{E}(n) \hookrightarrow \mathrm{Aff}(n)$ is a morphism of split extensions. If M_1^n and M_2^n are homotopy equivalent compact flat Riemannian manifolds, with fundamental groups Γ_1, $\Gamma_2 < \mathrm{E}(n)$ then Bieberbach shows that Γ_1 and Γ_2 are conjugate subgroups of $\mathrm{Aff}(n)$.

4.2. *Mostow rigidity*

The first version of Mostow's rigidity theorem appeared in [135]. Recall that a *hyperbolic manifold* is a complete Riemannian manifold whose sectional curvature is everywhere -1.

THEOREM 4.2 (Mostow rigidity theorem for hyperbolic manifolds). *If $f : M \to N$ is a homotopy equivalence between closed hyperbolic manifolds of dimension three or more then f is homotopic to an isometry.*

Mostow rigidity does not require constant curvature [136]. A *locally symmetric space* is a double coset space $\Gamma\backslash G/K$, where G is a semisimple Lie group, K is a maximal compact subgroup, and Γ is a discrete subgroup of G which acts on G/K as a group of covering transformations. A left-invariant Riemannian metric on G which is right K-invariant descends to a Riemannian metric on the double coset space.

THEOREM 4.3 (Mostow rigidity theorem for locally symmetric spaces). *Let M and N be compact locally symmetric spaces of non-positive sectional curvature. If the universal covering space $\widetilde{M}$ cannot be written as a metric product $M_1 \times M_2$, where one of the factors is of dimension one or two, then any homotopy equivalence $f : M \to N$ is homotopic to a diffeomorphism g, which may be assumed to be an isometry after the Riemannian metric on M is renormalized.*

Precursors to Mostow's rigidity theorem established rigidity for particular discrete subgroups in linear groups [155] or concluded that a discrete subgroup of a linear group is locally rigid: any continuous deformation carries Γ to a conjugate subgroup of G, as in the work of Weil [178–180]. The main construction of these papers of Weil leads to a description of the Zariski tangent space of the deformation space for a discrete subgroup of a linear group [126].

Mostow's rigidity theorem has been extremely influential, leading notably to the superrigidity and arithmeticity results of Margulis [129,130,187] and the analysis of Riemannian manifolds of nonpositive curvature and high rank [7,8].

5. Algebraic and topological foundations

Topological rigidity theorems are customarily proved in two or three main stages, passing through K-theory to a study of the structure set in topological surgery. A vanishing theorem for the Whitehead group of $\pi_1(M^n)$ is usually the first sign of progress. Although conventional wisdom claims that an idea which computes K-theory or proves a K-theoretic version of the Novikov Conjecture should eventually succeed in surgery computations, the gap between K- and L-theory arguments can be hard to bridge.

5.1. *K-theory obstructions*

The first obstruction to deforming a homotopy equivalence $f : M^n \to N^n$ to a homeomorphism is the obstruction to simple homotopy equivalence, i.e., the Whitehead torsion $\tau(f) \in \mathrm{Wh}(\pi_1(M))$.

The classes of manifolds for which rigidity theorems have been proved to date have much stronger vanishing properties than $\mathrm{Wh}(\pi_1(M)) \cong 0$. These are associated to a sequence of "Whitehead groups" $\mathrm{Wh}_i(\mathbb{Z}\Gamma)$ such that $\mathrm{Wh}_0(\mathbb{Z}\Gamma) \cong \widetilde{K}_0(\mathbb{Z}\Gamma)$, $\mathrm{Wh}_1(\mathbb{Z}\Gamma) \cong \mathrm{Wh}(\Gamma)$, and $\mathrm{Wh}_2(\mathbb{Z}\Gamma)$ agrees with the group defined by Hatcher–Wagoner in their study of pseudoisotopy.

The Loday assembly map for K-theory [125,170],

$$A : H_*\big(BG; \mathbb{K}(\mathbb{Z})\big) \to K_*(\mathbb{Z}G),$$

is part of the exact sequence in homotopy groups for a fibration of spectra

$$\mathbb{H}\big(BG;\ \mathbb{K}(\mathbb{Z})\big) \to \mathbb{K}(\mathbb{Z}G) \to \mathbb{W}\mathrm{h}^{\mathbb{Z}}(G)$$

in which $\pi_i(\mathbb{W}\mathrm{h}^{\mathbb{Z}}(G))$ agrees with the algebraically defined Whitehead groups $\mathrm{Wh}_i(\mathbb{Z}G)$ for $i = 0, 1, 2$ and may be taken as the definition of the Whitehead groups more generally. The Whitehead space thus measures the difference between the algebraic K-theory of G and a homology theory; this difference comes down to the possible failure of the excision axiom (Mayer–Vietoris sequences) for K-theory and is involved in splitting problems (Section 6).

K-theoretic versions of the Novikov Conjecture concern the assembly map A described above [90, pp. 32–33]. One of the most broadly applicable results in the realm of the Novikov and Borel Conjectures was proved in [14] using Waldhausen's A-theory [171], cyclic homology, and the authors' cyclotomic trace:

THEOREM 5.1 (Bökstedt–Hsiang–Madsen). *If Γ is a group such that $H_i(B\Gamma)$ is finitely generated for all i then*

$$A : H_*\big(B\Gamma; \mathbb{K}(\mathbb{Z})\big) \otimes \mathbb{Q} \to K_*(\mathbb{Z}\Gamma) \otimes \mathbb{Q}$$

is a split injection.

Conjectures discussed in Section 10 are posed for pseudoisotopy and A-theory as well as K-theory.

5.2. *Surgery theory*

Surgery on topological manifolds was a relatively late addition to a project which began in the smooth category and was shaped by smoothing and triangulation questions as well as classification problems. Although the topological theory is more complicated to establish than the smooth and PL versions, it has some formal properties which are particularly useful in rigidity arguments.

DEFINITION 5.1. The *structure set* of a compact manifold with boundary, M, is denoted $\mathcal{S}^{\mathrm{TOP}}(M, \partial M)$ or $\mathcal{S}(M, \partial M)$ and is defined to be the set of equivalence classes of pairs (N, f) where N is a compact manifold with boundary and $f:(N, \partial N) \to (M, \partial M)$ is a homotopy equivalence which restricts to a homeomorphism of boundaries; two pairs (N_1, f_1), (N_2, f_2) are equivalent if and only if there is a homeomorphism $h: N_1 \to N_2$ such that $f_2 \circ h$ is homotopic to f_1.

The Generalized Borel Conjecture (Conjecture 1.2) is conveniently phrased this way: If M is a compact aspherical manifold then $|\mathcal{S}^{\mathrm{TOP}}(M, \partial M)| = 1$. Structure sets for the categories of differentiable and piecewise linear manifolds are also defined and indicated by different decorations: $\mathcal{S}^{\mathrm{DIFF}}$, $\mathcal{S}^{\mathrm{PL}}$. In these terms, Milnor's celebrated theorem [132] reads as $|\mathcal{S}^{\mathrm{DIFF}}(S^7)| \neq 1$.

The Loday assembly map discussed above has a counterpart in surgery, introduced by Quinn. To a space X there is associated a spectrum $\mathbb{L}_\bullet(X)$ such that the homotopy groups of this spectrum are Wall's L-groups, $\pi_i(\mathbb{L}_\bullet(X)) \cong L_i(\mathbb{Z}[\pi_1(X)])$, and with a homotopy equivalence

$$\mathbb{L}_\bullet(\text{pt.})_0 \simeq L_0(\mathbb{Z}) \times G/TOP,$$

where G/TOP is the classifying space for topological reductions of stable spherical bundles. The $L_0(\mathbb{Z})$ factor in the 0-component of the surgery spectrum has become a matter of great interest [19], but to produce the long exact sequence of surgery we work with the 1-connective cover of $\mathbb{L}_\bullet(\text{pt.})$, denoted by $\mathbb{L}_\bullet$, which has $(\mathbb{L}_\bullet)_0 \simeq G/TOP$. The assembly map in surgery, $A : \mathbb{H}(X;\ \mathbb{L}_\bullet) \to \mathbb{L}_\bullet(X)$, is a map of spectra which appears in a cofibration sequence

$$\mathbb{H}(X;\ \mathbb{L}_\bullet) \to \mathbb{L}_\bullet(X) \to \mathbb{S}_\bullet(X). \tag{5.1}$$

There are versions of this construction for L-groups with different decorations, but since we work with manifolds whose Whitehead groups vanish we may ignore the distinction between surgery up to homotopy equivalence (L^h_*) and surgery up to simple homotopy equivalence (L^s_*).

If M^n is a closed n-dimensional manifold then the long exact sequence of homotopy groups in Equation (5.1) is identified by Poincaré duality with the surgery exact sequence of Browder–Sullivan–Wall [175, Chapter 10]. In particular, $\mathbb{S}_i(M) = \pi_i(\mathbb{S}_\bullet(M))$ may sometimes be identified with the topological structure set:

$$\mathcal{S}^{\mathrm{TOP}}\big(M^n \times D^k,\ \partial\big) = \mathbb{S}_{n+k+1}(M) \quad (k \geqslant 0).$$

The fact that the long exact sequence for topological surgery on a manifold is an exact sequence of Abelian groups is a considerable aid in computations. The classification argument for classifying tori [109,174,175] is an example of the difficulties imposed in categories other than TOP by their less robust surgery sequences.

5.3. *Siebenmann periodicity*

One of the important features of surgery in the topological category is a periodicity result due to Siebenmann [122, pp. 277–283] with corrections by Nicas [138].

THEOREM 5.2 (Siebenmann periodicity). *If* $(M^n, \partial M)$ *is a manifold with boundary and if* $n \geqslant 6$ *then there is an injection* $\mathcal{S}^{\mathrm{TOP}}(M^n) \to \mathcal{S}^{\mathrm{TOP}}(M^n \times D^4,\ \partial)$, *which is an isomorphism if* $\partial M \neq \emptyset$.

5.4. *Induction theorems*

The induction theorems needed here are modeled on the Frobenius reciprocity theorem for group representations. Dress [42] proved such theorems for the K- and L-groups and from homotopy theory we obtain similar results for the homological term in the long exact sequence of surgery theory [1, Chapter 4]. Induction methods were crucial in the spherical space form problem [166,127], and may be most familiar to topologists in that setting. It is important to note that versions of these theorems are valid for an infinite group Γ, provided we have a good finite quotient group of Γ.

The sketch below largely follows Nicas' memoir [138], which should be consulted for details.

Let M be a connected manifold with universal covering space $p: \widetilde{M} \to M$. Choose a basepoint $\widetilde{x}_0$ of $\widetilde{M}$, thus determining an action of $\Gamma = \pi_1(M, p(\widetilde{x}_0))$ on $\widetilde{M}$. To each subgroup $\Delta < \Gamma$ we associate the covering space $p_\Delta : \Delta\backslash\widetilde{M} \to M$. The inclusion $\Delta \hookrightarrow \Gamma$ and covering projection p_Δ determine transfer homomorphisms in K-groups, L-groups, and $H_*(-; \mathbb{L})$, which induce a transfer in the topological structure set.

Some transfer homomorphisms admit nice geometric descriptions associated to geometric interpretations of the groups involved. In particular, the transfer on $\mathcal{S}(M^n, \partial M)$ is simply pullback: If $[f] = [(V, \partial V) \to (M, \partial M)]$ is an the equivalence class of a homotopy equivalence of manifolds, restricting to a homeomorphism on the boundary, then $p_\Delta^*([f])$ is the equivalence class of the pullback or lift of f to $(V_\Delta, \partial(V_\Delta)) \to (M_\Delta, \partial(M_\Delta))$.

The K- and L-groups are covariant functors, where morphisms associated to a group homomorphism $\phi: G \to H$ are usually defined by tensoring an RG-module with RH, which is viewed as an (RG, RH)-bimodule. If ϕ is the inclusion of a finite-index subgroup then ϕ_* is modeled on the construction of induced representations, while the transfer morphism ϕ^* is modeled on the restriction operation in representation theory. Nicas [138] must work with the kernel and cokernel of the assembly map in the surgery exact sequence in order to draw conclusions about the topological structure set, whose group structure is essential for the argument.

Our goal is a test for triviality of elements x of the structure set $\mathcal{S}(M)$ of a manifold M. The assay for triviality is a collection of transferred images $p_\Delta^*(x) \in \mathcal{S}(M_\Delta)$, running over a family of finite-index subgroups Δ of $\pi_1(M)$: if each transferred element is trivial then we hope to conclude that x is trivial, i.e., represented by a homeomorphism. The family of subgroups used in the test is indexed by a collection of subgroups of a finite quotient group of $\pi_1(M)$.

Let C_k denote the cyclic group of order k and let C denote the infinite cyclic group.

DEFINITION 5.2. Let p be a prime number.

(a) A finite group G is *p-elementary* if it is the direct product of a cyclic group C_k and a p-group P, $G = C_k \times P$. G is *elementary* if it is p-elementary for some prime p.

(b) A finite group G is *p-hyperelementary* if it is a semidirect product of a cyclic group and a p-group, $G = C_k \times_\alpha P$. G is *hyperelementary* if it is p-hyperelementary for some prime p.

The next result is a weak version of [138, Proposition 6.2.9].

THEOREM 5.3 (Nicas). *Let M^n be a compact, connected manifold, where $n \geqslant 5$, let $\Gamma = \pi_1(M)$, and let $\psi : \Gamma \to Q$ be an epimorphism onto a finite group.*

Let $\mathcal{A}$ be the set of 2-hyperelementary subgroups of Q, let $\mathcal{B}$ be the set of p-elementary subgroups of Q where p runs through all the odd prime numbers, and let $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$.

Suppose that $x \in \mathcal{S}(M, \partial M)$ is an element of the topological structure set of $(M, \partial M)$. If $p^*_{\psi^{-1}(H)}(x) = 0$ for every $H \in \mathcal{C}$ then $x = 0$.

Nicas' result depends upon the work of Dress [42], which describes K- and L-groups as modules over functors with readily established induction properties, following the approach of J.A. Green to induced representations [96]. The next theorem is a statement of a weakened form of Dress' results adapted for our applications.

THEOREM 5.4 (Dress). *Let M^n be a compact, connected manifold, where $n \geqslant 5$, let $\Gamma = \pi_1(M)$, and let $\psi : \Gamma \to Q$ be an epimorphism onto a finite group. Let $\mathcal{D}$ be the set of elementary subgroups of Q.*

If $x \in \widetilde{K}_0(\mathbb{Z}\Gamma)$ then $x = 0$ if and only if the transfer $p^*_{\psi^{-1}(H)}(x) = 0$ for every $H \in \mathcal{D}$.

If $y \in \widetilde{\mathrm{Wh}}(\Gamma)$ then $y = 0$ if and only if the transfer $p^*_{\psi^{-1}(H)}(y) = 0$ for every $H \in \mathcal{D}$.

5.5. *Pseudoisotopy*

Pseudoisotopy is a tool for studying manifolds through their homeomorphism groups.

DEFINITION 5.3. A *topological pseudoisotopy* on a manifold M is a homeomorphism $h : M \times I \to M \times I$ such that $H|M \times \{0\}$ is inclusion. The space of all topological pseudoisotopies on M is denoted by $P(M)$.

K-theory and pseudoisotopy were shown to be linked by Hatcher and Wagoner [105]. Deeper connections emerged with the study of the stabilization operation $P(M) \to P(M \times I)$ obtained by forming the product $h \times \mathrm{Id}_I$. The space of *stable pseudoisotopies*, $\mathcal{P}(M)$, is the direct limit of the $P(M \times I^k)$ under stabilization. This space and its counterpart for diffeomorphisms of smooth manifolds have been shown by Hatcher to be Ω-spectra [104].

Waldhausen's A-theory is the context in which K-theory and pseudoisotopy seem to make their nearest approach [171,172]. We are unable to say much about this subject here, but rigidity theorems and conjectures cited below will involve pseudoisotopy.

6. Splitting and fibering problems

Dissection arguments are a mainstay of classification results in topology, with classical roots in the study of a Riemann surface by cutting it open along curves. Waldhausen proved topological rigidity by cut and paste methods for a large family of aspherical 3-manifolds [169]:

THEOREM 6.1 (Waldhausen). *Let M^3 and N^3 be compact, connected, sufficiently large, irreducible* 3*-manifolds. If $f:(M, \partial M) \to (N, \partial N)$ is a homotopy equivalence which restricts to a homeomorphism of boundaries then f is homotopic to a homeomorphism.*

Classification or rigidity arguments which induce over a sequence of cut-and-paste constructions usually begin with a description of the target manifold N via manifolds with boundary obtained by cutting along codimension-one, two-sided submanifolds; we seek to transport this structure in N to the source manifold M, and then to solve a relative version of the homeomorphism problem in the induction step, as in the conclusion and proof of Waldhausen's theorem above. The step in which a dissection of N is transported to M involves more than transversality, since at minimum we seek to deform a homotopy equivalence of single spaces $M^n \to N^n$ to a homotopy equivalence of pairs $(M^n, V^{n-1}) \to (N^n, W^{n-1})$.

6.1. *Splitting problems in K-theory*

Let N^n be a manifold with a a codimension-one submanifold W. This submanifold is said to be two-sided if it has an open neighborhood U such that $(U, W) \cong (W \times (-1, 1), W \times \{0\})$.

Let N^n be a manifold with a connected, two-sided submanifold W^{n-1}. A *splitting problem* is a homotopy equivalence $f: M^n \to N^n$; a solution to the splitting problem is a map g which is homotopic to f, transverse to W^{n-1} (so that $V = g^{-1}(W)$ is a two-sided, codimension-one submanifold of M), and so that $g:(M^n, V^{n-1}) \to (N^n, W^{n-1})$ is a homotopy equivalence of pairs. One often says that a homotopy equivalence f *splits* along W if it may be deformed to a map g as above.

Obstructions for splitting problems arise in both K-theory and surgery, and may be interpreted as exotic terms in generalized Mayer–Vietoris sequences for those theories. These exotic terms are K-theoretic groups in their own right, called Nil groups; see [9, 170] for more information.

Waldhausen followed his results on sufficiently large 3-manifolds, which imply the vanishing of Whitehead groups for fundamental groups of such manifolds, with an algebraic study of K-theoretic splitting obstructions [170]. This study analyzes Whitehead groups for HNN extensions and free products with amalgamations, and we describe those results below: a finite induction with these methods allows Waldhausen to deduce a vanishing theorem which may be posed in terms of graphs of groups [158].

A ring R is regular coherent if every finitely presented right R-module admits a finite resolution by finitely generated projective R-modules, and R is regular Noetherian if it is regular coherent and every finitely generated R-module is finitely presented. A group G is regular coherent if for every regular Noetherian ring R the group ring RG is regular coherent. (For example, finitely generated free Abelian groups have regular Noetherian group rings, while surface groups are regular coherent.)

Let $\mathcal{C}$ be the smallest class of groups containing the trivial group, closed under the construction of finite graphs of groups with with regular coherent edge groups, and closed under filtering direct limits.

THEOREM 6.2. *If R is regular Noetherian and G is a group in $\mathcal{C}$ then* $\mathbb{W}\mathrm{h}^R(G)$ *is contractible.*

COROLLARY 6.1. *If G is the fundamental group of a compact irreducible* 3*-manifold which is sufficiently large then* $\mathrm{Wh}_i(\mathbb{Z}G) = 0$ *for all* $i \geqslant 0$.

6.2. *Fibering obstructions*

Bass, Heller, and Swan [9] anticipated Waldhausen's splitting theorem in K-theory in the important special case of a group of the form $G = H \times C_\infty$, where C_∞ denotes the infinite cyclic group. This result and its extension from direct to semidirect products [51] are part of the analysis of a geometric problem related to splitting:

PROBLEM 6.1 (*The Fibering Problem*). When is a manifold M^n homeomorphic to the total space of a bundle over a circle with closed manifold fiber?

Irreducible 3-manifolds which fiber over the circle are well understood [162,106], and those results motivated the study of fibering problems in higher dimensions.

The high-dimensional fibering problem was solved by Farrell in his 1967 thesis [47], which exhibited complete obstructions in K-theory (see also [160]). The argument uses h-cobordism methods to produce a product structure on an infinite cyclic covering space. The fibering theorem was extended to the boundary dimensions [181] after Freedman's seminal work on 4-manifolds [91,92]. (Recall that restrictions on the fundamental group are currently required for 4-dimensional surgery: see [92].)

6.3. *Splitting problems in surgery*

The surgical aspect of splitting problems was examined by Wall [175, Chapters 11 and 12] and Cappell [21–23], following earlier work of Shaneson [159] and Wall on L-groups of products with an infinite cyclic group.

In [23] Cappell proved a splitting theorem under the assumptions that K-theory obstructions vanish and that $\pi_1(V^{n-1})$ is a square-root closed subgroup of $\pi_1(M^n)$. (A subgroup $H < G$ is *square-root closed* if $g \in G$ and $g^2 \in H$ imply $g \in H$.) This result and extensions of it in the algebraic theory of surgery (see Ranicki's work) have been used to prove cases of both the Novikov Conjecture and Borel Conjecture ([139] is one of several examples).

Cappell's announcements [21,22] define counterparts to the Nil groups of K-theory, called UNil groups.

6.4. *Nil and UNil groups*

The UNil groups, like the Nil's, appear as exotic terms in Mayer–Vietoris sequences, and if they are nontrivial for splitting data (M^n, V^{n-1}) then one expects the structure set of M^n to contain more than one element, and indeed to be large [22].

Cappell shows that the UNil groups are sometimes nontrivial by analyzing the free product of two cyclic groups of order two, $D_\infty = C_2 * C_2$. Since this infinite dihedral group is a crystallographic group, one should not be surprised to encounter difficulties as rigidity investigations move from manifolds to orbifolds (Section 10). The Borel Conjecture predicts that all UNil groups vanish for a torsion-free fundamental group and no evidence is presently known against this claim.

Some order is emerging in the picture of Nil, UNil, and related groups in K-theory [33, 48,49], but more work is needed.

7. Controlled topology

The next few paragraphs are a guide to parts of the rapidly growing literature on controlled topology, emphasizing topics associated with rigidity theorems. See the chapter on homology manifolds in this volume for more detail and a broader perspective.

Connell and Hollingsworth [30] published in the late 1960s a description of Whitehead torsion in terms of *geometric modules*: such an object is a based free module M over a ring, together with a function from the basis to a metric space (X, d). This function gives a notion of support in X for an element of M, and we say that a module element is ε-small if its support has diameter less than ε. We get a similar notion of the diameter of an endomorphism of M. Connell and Hollingsworth reduced the topological invariance of Whitehead torsion to certain problems about geometric modules and morphisms, but were unable to resolve these.

In the late 1970s Ferry [84,85] and Chapman and Ferry [27] proved approximation theorems, including results that show that a sufficiently controlled homotopy equivalence must be a simple equivalence. The resulting Thin h-Cobordism Theorem (see Theorem 7.1) was at the heart of the rigidity program for flat manifolds. Chapman later formulated counterparts of the finiteness obstruction and Whitehead torsion for controlled maps and spaces in response to Quinn's work [25,26].

Also in the late 1970s, Farrell and Hsiang began work on flat manifolds other than the torus with sharp results on K-theory and partial results on surgery [52], proving topological rigidity of flat Riemannian manifolds with holonomy groups of odd order.

Quinn's 1979 paper [143] studied ends of spaces and maps with controlled methods; this paper also revived the geometric modules of Connell and Hollingsworth and established some of their conjectures (see also [146,93] and [50, Lectures 9–11]). Quinn continued his influential work on ends in [144,145] and Quinn's student Yamasaki has worked on K-theory of crystallographic groups as well as foundations [150,186].

The completion of the proof of topological rigidity for flat Riemannian manifolds [56] by Farrell and Hsiang covered almost flat manifolds as well. The argument depended upon fiberwise control results established in [56] and implicit in [143]. These fiberwise results are precursors of the foliated control theorems employed by Farrell and Jones in their work on manifolds of nonpositive curvature [58,60–62,64,65,67,69].

Another vein of controlled argument emphasizes bounded structures and is represented by [2,3,141], and a number of other papers. The bounded theory has better categorical properties than ε-controlled algebra.

7.1. *Control theorems*

Recall that an h-cobordism is a manifold W^{n+1} whose boundary is partitioned as a disjoint union $\partial W = M_0^n \coprod M_1^n$, where both inclusions $i_j : M_j \hookrightarrow W$ are homotopy equivalences. (We often call M_0 the *base* of the h-cobordism.) It follows that there exist deformation retractions $r_j : W \times I \to W$ such that $r_j(w, 0) = w$ for every $w \in W$, $r_j(w, 1) \in M_j$ for every $w \in W$, and $r_j(x, t) = x$ for every $x \in M_j$ and every $t \in I$.

DEFINITION 7.1. Let (W^{n+1}, M_0^n, M_1^n) be an h-cobordism and let d be a metric on M_0. W is *ε-controlled* provided there exist deformation retractions r_0 and r_1 of W to M_0 and M_1, as above, such that for every $w \in W$ the paths ρ_w and σ_w defined by

$$\rho_w(t) = r_0\big(r_0(w, t),\ 1\big) \quad (t \in I)$$

and

$$\sigma_w(t) = r_0\big(r_1(w, t),\ 1\big) \quad (t \in I)$$

have diameter less than ε in M_0.

The controlled vanishing theorem for Whitehead torsion proved by Ferry in 1977 has been extremely influential [84]:

THEOREM 7.1 (Thin h-cobordism theorem). *For each closed Riemannian manifold M^n of dimension $n > 4$ there exists a real number $\varepsilon > 0$ such that any ε-controlled h-cobordism W with base M has trivial Whitehead torsion.*

A similar result was proved by Chapman and Ferry for homeomorphisms [27,85].

DEFINITION 7.2. Let M_0 be a Riemannian manifold with metric d, and let ε be a positive real number. Suppose that $f : M_1 \to M_0$ is a homotopy equivalence. We say that f is an *ε-equivalence* if there exist a homotopy inverse $g : M_0 \to M_1$, a homotopy $F : M_0 \times I \to M_0$ from $f \circ g$ to Id_{M_0}, and a homotopy $G : M_1 \times I \to M_1$ from $g \circ f$ to Id_{M_1} so that for every $x \in M_0$ the path ρ_x defined by

$$\rho_x(t) = F(x, t) \quad (t \in I)$$

has diameter less than ε in M_0 and for every $y \in M_1$ the path σ_y defined by

$$\sigma_y(t) = f\big(G(y,t)\big) \quad (t \in I)$$

has diameter less than ε in M_1.

THEOREM 7.2 (Chapman–Ferry α-approximation theorem). *Let N^n be a compact Riemannian manifold of dimension $n \geqslant 5$. There is an $\varepsilon > 0$ such that any ε-equivalence $f:(M,\partial M) \to (N,\partial N)$ which restricts to a homeomorphism of boundaries is homotopic to a homeomorphism.*

The "α" of the title denotes an open cover of N; the full strength of the theorem works with open covers rather than metric conditions, and applies to noncompact spaces. This style of controlled vanishing or rigidity result has also been established for the projective class group $\widetilde{K}_0$ and the structure set of topological surgery [56].

7.2. *Properties of flat manifolds*

The rigidity argument presented in Section 7.3 was largely motivated by the theorem of Epstein and Shub [46] which asserts that every closed flat Riemannian manifold supports an expanding diffeomorphism. This suggests an attack on invariants which can be represented by geometric modules or bounded morphisms of geometric modules: transfer an element x through a diffeomorphism h which is expanding enough that the diameter of $h^*(x)$ is so small that a control theorem implies $h^*(x) = 0$.

Unfortunately, there is no reason for transfer by an expanding map to induce an isomorphism in K-theory or surgery. A more elaborate argument is required, which uses Frobenius induction and multiple covering spaces of flat manifolds, as well as a vanishing result for sufficiently expansive covering projections.

Flat manifolds and orbifolds of dimension n are quotients of $\mathbb{R}^n$ by discrete subgroups $\Gamma < \mathrm{E}(n)$, where $\mathrm{E}(n)$ is the group of Euclidean isometries (Section 4.1). In general, such a Γ is called a *crystallographic group*; if Γ is torsion-free and cocompact in $\mathrm{E}(n)$ then Γ is called a *Bieberbach group*.

Bieberbach showed [185, Chapter 3] that the group extension decomposition of the group $\mathrm{E}(n)$ of isometries of $\mathbb{R}^n$ exhibited in Equation (4.1) is preserved in cocompact discrete subgroups $\Gamma < \mathrm{E}(n)$: Γ has a maximal normal free Abelian subgroup A, which consists of translations, and Γ is an extension

$$1 \to A \to \Gamma \to G \to 1, \tag{7.1}$$

where $A \cong C^n$ is a free Abelian group and G is a finite group.

The finite group G is known as the *holonomy group* of Γ, and although Γ is imbedded in the split extension of Equation (4.1), Equation (7.1) is not usually a split extension. Auslander and Kuranishi showed that every finite group appears as the holonomy group of some Bieberbach group [185, Theorem 3.4.8].

Every abstract group Γ possessing a normal free Abelian subgroup of finite index is realized as a crystallographic group, so rigidity theorems proved for flat manifolds by geometric methods may be recast as theorems concerning a certain class of abstract groups.

7.3. *Topological rigidity of flat manifolds*

The vanishing theorem of Farrell and Hsiang for K-theory obstructions [52, Theorem 3.1] was the first general result on rigidity questions for flat manifolds, and serves as a starting point for later arguments.

THEOREM 7.3 (Farrell–Hsiang). *If Γ is a Bieberbach group then* $\mathrm{Wh}(\Gamma) = 0$ *and* $\widetilde{K}_0(\mathbb{Z}\Gamma) = 0$.

The outline of the proof begins with the observation that every crystallographic group has a non-elementary finite quotient group, so we may apply Theorem 5.4. One might expect an induction over the order of the holonomy group of Γ to yield the conclusion, but the existence of expanding maps on flat manifolds [46] is one aspect of a fact which is awkward for the induction argument: A flat manifold may be a covering space of itself. We say that a self-map is infinitesimally s-expansive if its differential satisfies $||Df(v)|| = s||v||$ for every tangent vector v.

Farrell and Hsiang are obliged to prove a structural result for Bieberbach groups Γ [52, Theorem 1.1] and a quantitative vanishing result for expansive covering projections [52, Theorem 2.3], which shows that if an expansive self-map of a closed flat manifold is infinitesimally s-expansive for sufficiently large s then the associated transfer annihilates the Whitehead and projective class groups.

The main argument is a double induction, running over both the rank of the translation subgroup of Γ and the order of the holonomy group. (Note that the seed of the induction is the free Abelian case, $\Gamma \cong C^n$, which is settled by the Bass–Heller–Swan theorem [9].) The induction step uses a non-elementary finite quotient G of Γ, which is chosen so that if $H < G$ is covered by a subgroup $\Gamma_H < \Gamma$ which is isomorphic to Γ then the associated covering projection is infinitesimally s-expansive for s large enough to annihilate the K-groups. We have arranged matters so that for each elementary subgroup H of G, either (a) we reduce to a case already settled by the induction hypothesis or (b) the subgroup of Γ lying over H is isomorphic to Γ and we are in the infinitesimally s-expansive case for large s. Both alternatives are won cases, so Theorem 5.4 finishes the argument.

The analogous argument for surgery theory was less satisfactory in [52] because of difficulties with splitting obstructions. Improvements in [57] give a rigidity theorem.

THEOREM 7.4 (Farrell–Hsiang). *Let M^n be a closed aspherical manifold whose fundamental group is virtually nilpotent and let E^{n+k} be the total space of a D^k-bundle over M^n. If $n+k > 4$ then $\mathcal{S}(E^{n+k}) = 0$; in particular, if $n > 4$ then M is topologically rigid.*

The scheme adopted for this argument, which covers almost flat manifolds (Definition 3.2) as well as flat manifolds, was called a “fibering apparatus”, and has been adopted in later papers by Lee and Raymond. The main feature of the scheme for the present account is the control theorem proved in [56], in which diameters are measured in the base of a fiber bundle whose fibers are already known to have good rigidity properties. The notion that control could be maintained in some directions and relaxed in others was the key to rigidity results for Riemannian manifolds of nonpositive curvature.

7.4. *Properties of negatively curved manifolds*

Two features seen in a compact Riemannian manifold M^n of nonpositive sectional curvature have been exploited effectively by topologists. The first and most readily generalized of these is the compactification of the universal covering space $\widetilde{M}^n$ by adjoining an $(n-1)$-sphere at infinity defined by geodesic rays [43]. This boundary sphere will be denoted by $\partial_\infty M$ below.

The visibility sphere V_x at a point $x \in \widetilde{M}^n$ is the unit sphere in $T_x\widetilde{M}$, which we identify with the set of unit-speed geodesic rays based at x. With the Cartan–Hadamard theorem and a radial reparametrization of the exponential map based at x, we identify $\widetilde{M}$ with the interior of the disk D^n and we identify the visibility sphere V_x with the boundary of D^n, compactifying $\widetilde{M}$. This compactification, denoted here by $\mathrm{VC}(\widetilde{M})$ appears heavily dependent upon x and to make it more natural we introduce an identification of visibility spheres $V_x \cong V_y$. Two geodesic rays ρ, σ are asymptotes, or asymptotically equivalent, if there is a constant C such that for all $t \geqslant 0$ $d(\rho(t), \sigma(t)) \leqslant C$. If M is a complete Riemannian manifold of nonpositive sectional curvature then this equivalence relation on elements of visibility spheres at different basepoints makes the compactification of $\widetilde{M}$ by the sphere at infinity naturally enough that the isometry group $\mathrm{Isom}(\widetilde{M})$ acts by homeomorphisms upon $\mathrm{VC}(\widetilde{M})$.

The compactification of $\widetilde{M}^n$ to a disk by $\partial_\infty M$ and emulations of this compactification, as for word hyperbolic groups [98], play an important role in most approaches to the Novikov Conjecture for nonpositively curved manifolds and related groups. (See [88,89] for more information.) We want to use the fact that if $\sigma(t)$ $(-\infty < t < \infty)$ is a geodesic in $\widetilde{M}^n$ then this curve has well-defined endpoints in the sphere at infinity, $\sigma(-\infty), \sigma(\infty) \in \partial_\infty \widetilde{M}^n$

The second geometric feature to discuss is the geodesic flow Let SM denote the subbundle of the tangent bundle TM consisting of tangent vectors of length 1, so that SM has fiber S^{n-1}. Let $p: TM \to M$ be the projection in the tangent bundle and let $q = p|_{SM}$ be the projection in the unit tangent sphere bundle. Define a vector field Σ on TM by $\Sigma_v = (v, 0) \in T_vTM \cong T_{p(v)}M \oplus T_{p(v)}M$, where the first summand is horizontal and the second vertical in T_vTM. Σ is known as the *geodesic spray* and restricts to a vector field of length 1 everywhere on SM. The *geodesic flow* is the flow ϕ_t on SM of the vector field $\Sigma|_{SM}$.

LEMMA 7.1. *Let ϕ_t be the geodesic flow on the unit tangent sphere bundle SM of a complete Riemannian manifold.*

(b) *ϕ_t is a complete flow, i.e., the flow line through any initial point exists for all t.*

(b) *Integral curves of ϕ_t cover geodesics in the following sense: For any $v \in SM$ $t \mapsto q \circ \phi_t(v)$ is the geodesic in M through $q(v)$ with initial vector v.*

Anosov's study of the geodesic flow [4] showed that the following features of the flow imply many of its other properties, and tend to rigidify it among flows; the fact that the geodesic flow does have these properties essentially go back to Hadamard. Any smooth flow which satisfies (a)–(c) below is now called an *Anosov flow*.

LEMMA 7.2. *Let SM be the unit tangent sphere bundle of a complete Riemannian manifold of sectional curvature* $-b^2 \leqslant K \leqslant -a^2 < 0$.

If ϕ_t *is the geodesic flow on SM then*:

(a) *The tangent bundle TSM splits continuously into a Whitney sum of three* ϕ_t*-invariant subbundles*, $TSM = E^0 \oplus E^u \oplus E^s$, *where*
(b) E^0 *is tangent to the geodesic flow, and*
(c) *There exist positive constants C and* λ *such that*

$$||D_x\phi_t|_{E^s_x}|| \leqslant Ce^{-\lambda t} \quad \textit{and} \quad ||D_x\phi_{-t}|_{E^u_x}|| \leqslant Ce^{-\lambda t}.$$

The subbundles E^s and E^u are the *stable subbundle* and the *unstable subbundle*, respectively. In the most familiar example of this construction, the hyperbolic plane $M^2 = H^2_{\mathbb{R}}$ is realized as the unit disk, $SM = M^2 \times S^1$, and the 3-dimensional tangent space T_vSM splits as a sum of three line bundles. If $x = q(v) \in M^2$ and $\sigma(t)$ is the unit-speed geodesic through x with initial tangent vector $\mathrm{d}\sigma(t)/\mathrm{d}t|_{t=0} = v$ then the stable and unstable line bundles may be described in terms of horocycles through x, orthogonal to v, and passing through $\sigma(\infty)$ (in the stable case) or $\sigma(-\infty)$ (in the unstable case). An illustrated account of this example may be found in [11, Chapter 3].

The geodesic flow is a well-studied object in dynamics and geometry, largely because of its structural stability and ergodic properties (on manifolds of finite volume) and the fact that its periodic orbits are smooth closed geodesics. We need the much more basic properties of the geodesic flow stated in Lemma 7.2.

7.5. *Topological rigidity of nonpositively curved manifolds*

While the Farrell–Hsiang program on flat and almost flat manifolds was underway W.C. Hsiang suggested that the distinctive metric features of the geodesic flow might be the foundation for rigidity arguments addressing negatively curved manifolds. A topological counterpart to Mostow rigidity for hyperbolic manifolds (Theorem 4.2) was particularly sought.

This ambition was realized by a sequence of papers of Farrell and Jones which began in 1986 [58] and continued for at least ten years. The new ingredient from controlled topology is a vanishing theorem in the mode of Ferry–Chapman–Quinn, but with control hypotheses adapted to the squeezing/expanding behavior of an Anosov flow. Since the geodesic flow lives on the unit tangent sphere bundle SM rather than on M, a device is needed to move representative elements of obstruction groups from M to SM (we have in mind, for example, the Whitehead group of $\pi_1(M)$ and the realization of each element of this group as the Whitehead torsion of an h-cobordism based on M).

The *asymptotic transfer* lifts a path $\alpha: I \to M$ to a path $v\alpha: I \to SM$ such that $q \circ (v\alpha) = \alpha$. The notion of asymptotic equivalence for geodesic rays discussed in Section 7.4 is visualized directly within the visibility compactification: any geodesic ray has a well-defined endpoint in $\partial_\infty \widetilde{M}$, and two geodesic rays are asymptotes if and only if their endpoints in the sphere at infinity agree. The asymptotic transfer will first be defined for paths in $\widetilde{M}$; because the construction is equivariant for the action of the isometry group, it descends to M.

For each $v \in SM$ let γ_v be the geodesic in $\widetilde{M}$ such that $\dot{\gamma}_v(0) = v$, where the overdot indicates derivative with respect to the time parameter along the geodesic. (Recall that the geodesic equation is a second order differential equation, so we are solving the initial value problem $\gamma_v(0) = q(v)$, $\dot{\gamma}_v(0) = v$.) For each $v \in S\widetilde{M}$ and each $x \in \widetilde{M}$ let $v(x) \in S_x M$ be the unique unit vector at x such that γ_v and $\gamma_{v(x)}$ are asymptotes. We define the asymptotic transfer $v\alpha$ of the path α by

$$v\alpha(t) = v\big(\alpha(t)\big) \quad (t \in I).$$

The most important property of the asymptotic construction is a squeezing statement: The geodesic flow on SM shrinks $v\alpha$ in all directions normal to the flow lines, so that the flow deforms $v\alpha$ arbitrarily close to some flow line as $t \to +\infty$, while the length of the deformed curve remains bounded. (This metric statement is based on the fact that $v\alpha$ is contained in a leaf of E^s.)

DEFINITION 7.3. A path ρ in the unit tangent sphere bundle SM is (β, ε)-controlled if there is another path σ in SM such that:

(a) The image of σ is contained in an arc of length β inside a flow line of the geodesic flow.

(b) $d(\rho(t), \sigma(t)) < \varepsilon$ for every $t \in I$.

LEMMA 7.3. *Given $\beta > 0$ and $\varepsilon > 0$ there exists $s_0 > 0$ such that for every smooth path α in M of length less than β and for every vector $v \in S_{\alpha(0)}M$ the composite with the geodesic flow $\phi_s(v\alpha)$ is $(\sqrt{2}\beta, \varepsilon)$-controlled for every $s \geqslant s_0$.*

This is one of the foliated control theorems established by Farrell and Jones, specialized from a theorem in [61].

THEOREM 7.5 (Foliated control for h-cobordisms). *Given M^n, a closed manifold of strictly negative sectional curvature, where $n > 2$, and a positive real number β, there exists $\varepsilon > 0$ such that every (β, ε)-controlled h-cobordism with base SM has trivial Whitehead torsion.*

The argument for this theorem follows the model of Ferry [84], after SM is equipped with a "long-thin cell structure" adapted to the dynamics of the geodesic flow. Using this foliated control theorem, an argument for the vanishing of $\mathrm{Wh}(\pi_1(M))$ may be built on the following outline:

(1) Represent $x \in \mathrm{Wh}(\pi_1(M))$ by a smooth h-cobordism W with base M.

(2) Take smooth deformation retractions of W onto its top and base and let β bound the arc lengths of the deformation paths or tracks for these homotopies (as measured in M).
(3) Let $\mathcal{W}$ be the total space of the pullback of $q: SM \to M$ to W (pull back over the retraction onto the base of the h-cobordism). $\mathcal{W}$ is now an h-cobordism, which we equip with top and base retractions whose tracks are asymptotic transfers of the tracks in M.
(4) Take ε from Theorem 7.5 for the data $\mathcal{W}$ and $\sqrt{2}$. Flow forward with the geodesic control long enough to make $(\mathcal{W}, SM)$ $(\sqrt{2}\beta, \varepsilon)$-controlled.
(5) Apply Theorem 7.5 to conclude that the asymptotically transferred h-cobordism has trivial Whitehead torsion.

A difficulty appears at the end of the argument: we do not know if asymptotic transfer induces an isomorphism in Whitehead groups, and in fact the known transfer theorems for algebraic K-theory don't allow us to draw such a conclusion because we are transferring through a fiber bundle projection where the spherical fiber has Euler characteristic 0 or 2, depending on n. Farrell and Jones modified the construction so that transfer is performed in a related, but noncompact, bundle whose fiber has Euler characteristic 1.

Arguments of this kind have led to a number of conclusions, including these from [77, 80].

THEOREM 7.6 (Rigidity of nonpositively curved manifolds). *Let M^n be a closed Riemannian manifold of nonpositive sectional curvature.*
(a) $\mathrm{Wh}(\pi_1(M)) = 0$.
(b) *If* $m + n \geqslant 5$ *then* $\mathcal{S}(M^n \times D^k, \partial) = 0$.

Subsequent developments have required modifications of the asymptotic transfer to a "focal transfer" which focuses on a point at finite distance from a lifted path rather than on a point at infinity. Noncompact manifolds have also been studied by Farrell and Jones with these methods, and some of this work remains in progress as this chapter is written. Their paper on isomorphism conjectures [78] considers orbifolds as well as manifolds (see Section 10) and refines our understanding of the role of geodesic flow. For many computations it seems likely that we can view elements of obstruction groups as concentrated along closed loops (or along periodic orbits for the geodesic flow in a negatively curved manifold).

Geometers and dynamicists are currently investigating rigidity properties of geodesic flows (to what degree does the dynamics of the flow determine the Riemannian metric?). We quote only one of these results. C.B. Croke, P. Eberlein, and B. Kleiner have established the following rigidity result [37].

THEOREM 7.7 (Croke, Eberlein, and Kleiner). *Let M and N be compact Riemannian manifolds of sectional curvature $K \leqslant 0$, such that M has dimension three or more and rank at least two. If there is a C^0 conjugacy F between the geodesic flows on the unit tangent sphere bundles of M and N then there exists an isometry $G: M \to N$ that induces the same isomorphism as F on fundamental groups.*

8. Exotic structures

8.1. *Manifold structures*

We now know that the full range of variations of geometric structures can be realized in aspherical manifolds, often with interesting side conditions.

Davis and Hausmann showed that there are closed aspherical manifolds with no smooth or PL structure [40]:

THEOREM 8.1 (Davis–Hausmann). (a) *For each $n \geqslant 13$ there exists an aspherical closed PL manifold of dimension n which does not have the homotopy type of a smooth manifold.*

(b) *For each $n \geqslant 8$ there exists an aspherical closed topological manifold of dimension n which is not homeomorphic to a closed PL manifold.*

Farrell and Jones have found exotic smooth structures on compact or noncompact manifolds of negative curvature. In the compact case [71] they take a connected sum of a hyperbolic manifold and an exotic sphere, but in the noncompact case a connected sum is known to be inadequate for the creation of nonstandard smooth structures. In [79] Farrell and Jones modify smooth structures with a noncompact counterpart to Dehn surgery, modifying a properly imbedded tube.

In both the compact and noncompact cases, Farrell and Jones show that some of the exotic smooth structures they produce can be equipped with a Riemannian metric of negative sectional curvature. This construction limits the possible scope of rigidity theorems or pinching theorems in differential geometry and also leads to some new results on harmonic maps.

8.2. *Exotic harmonic maps*

A *harmonic map* between smooth manifolds is a critical point for an energy functional which is roughly the mean square of the covariant derivative. These maps are solutions of nonlinear elliptic partial differential equations and have good existence and regularity properties in several different contexts. (An example which is essentially familiar to topologists is a harmonic map from a closed manifold to the circle S^1: such a map pulls back $d\theta$ to a harmonic 1-form in the sense of Hodge theory.)

Differential geometers had conjectured that a harmonic homotopy equivalence between closed manifolds of nonpositive curvature must always be a homeomorphism. Some of the evidence for such a claim is found in the role of harmonic map techniques in recent work on rigidity and arithmeticity done by K. Corlette, C. Simpson, M. Gromov and R. Schoen, and others. Farrell and Jones used the classic existence results of Eells and Sampson on harmonic maps to a nonpositively curved target and their constructions of exotic smooth structures to produce singular harmonic homotopy equivalences [83].

THEOREM 8.2. *Let $n \geqslant 11$ and suppose that M^n is a closed smooth manifold with a Riemannian metric g such that either (M, g) has negative sectional curvature or (M, g) is a flat torus. Then there exists a second Riemannian metric h on M and a harmonic homotopy equivalence $f : (M, h) \to (M, g)$ such that f is not injective.*

8.3. *Homology manifolds*

The ENR homology manifolds introduced by Bryant, Ferry, Mio, and Weinberger [19] suggest non-manifold variations on some of the questions raised above.

If M is a closed aspherical manifold with trivial Whitehead group and if the surgery assembly map $A : H_i(M; \mathbb{L}) \to L_i(\mathbb{Z}\pi_1(M))$ is an isomorphism for all i then M is rigid in the class of ENR homology manifolds: there is no exotic ENR homology manifold with the homotopy type of M, and every closed manifold homotopy equivalent to M is homeomorphic to M [86].

An ENR homology manifold version of the realization problem for Poincaré duality groups has also been suggested. There seems to be no progress to report as this chapter is written.

9. Poincaré duality groups

This notion, mentioned in Section 3.1, was introduced by Bieri [13] and Johnson and Wall [116]. The abbreviation "PD^n group" is often used in discussing Poincaré duality groups.

DEFINITION 9.1 (*Poincaré duality group*). A *Poincaré duality group* of formal dimension n is a countable group Γ of type FP such that

(a) $D = H^n(\Gamma; \mathbb{Z}\Gamma)$ is an infinite cyclic Abelian group and

(b) there is a class $z \in H_n(\Gamma; D)$ so that the cap product map

$$- \cap z : H^i(\Gamma; M) \to H_{n-i}(\Gamma; D \otimes M)$$

is an isomorphism for all Γ-modules M and all integers i.

See Definition 3.1 for the FP property, and see Brown [18, Section VIII.10] for more information on Poincaré duality groups. (The handbook survey [10] on group cohomology may also be useful.) Brown shows that if (a) and (b) hold Γ is necessarily of type FP [18, p. 222], but we include the extra hypothesis for clarity.

The following theorem combines work of Eilenberg–Ganea and Wall [18, Section VIII.7]. Here $\mathrm{cd}(\Gamma)$ denotes the cohomological dimension of the discrete group Γ.

THEOREM 9.1. *Let Γ be an arbitrary group and let $n = \max(3, \mathrm{cd}(\Gamma))$. Then there exists an n-dimensional $K(\Gamma, 1)$-complex Y. If Γ is finitely presented and of type FL (respectively FP) then Y can be taken to finite (respectively finitely dominated).*

Note that the fundamental group of a finitely dominated complex is necessarily finitely presented [173, Lemma 1.3], but that the homological finiteness properties FL and FP do not directly lead to a finite presentation for a group. The long-standing question raised by this situation was resolved by Bestvina and Brady [12]:

THEOREM 9.2 (Bestvina–Brady). *There are groups of type FP which are not finitely presented.*

Building upon the construction of Bestvina and Brady, M. Davis has announced examples of groups which satisfy Poincaré duality (conditions (a) and (b) in Definition 9.1) but which do not have finitely dominated classifying spaces, because these groups are not finitely presented.

The following question has been posed in different flavors at different times [176,177]. The recent developments summarized above oblige us to assume finite presentability.

PROBLEM 9.1 (*Realization problem for Poincaré duality groups*). Is every finitely presented Poincaré duality group the fundamental group of a closed aspherical manifold?

The attack on this question has produced good results in dimension 2, due to Eckmann and his collaborators [45,44], and some information on special classes of groups [6,115] for which induction arguments work well and group extensions are closely related to geometry. Some information has been obtained in dimension 3 [167,107], largely through splitting and fibering arguments.

Surgery theory has an existential aspect which gives a scheme for deciding whether a finite complex that satisfies Poincaré duality is homotopy equivalent to a manifold. The language of the structure set (Definition 5.1) extends to this situation and we seek to show that $|\mathcal{S}^{\mathrm{TOP}}(K(\Gamma, 1))| \neq 0$ if Γ is a Poincaré duality group with finite $K(\Gamma, 1)$. The most general forms of the Borel Conjecture have implications for Problem 9.1, but do not seem to be established except for groups we already know are realized by aspherical manifolds.

A characterization and approximate realization of Poincaré duality groups appears in [165]:

THEOREM 9.3. *Let Γ be a finitely presented group of finite cohomological dimension. Γ is a Poincaré duality group if and only if there exists a closed PL manifold M with fundamental group Γ and universal cover $\widetilde{M}$ such that $\widetilde{M}$ is homotopy equivalent to a finite complex.*

If Γ has a finite $K(\Gamma, 1)$-complex then the space $\widetilde{M}$ of the theorem may be taken to be homotopy equivalent to a sphere of arbitrarily high dimension. The question of whether or not a finitely presented Poincaré duality group is of type FL and so has a finite Eilenberg–MacLane space remains a considerable obstacle in the realization problem.

The number of known Poincaré duality groups grew dramatically in the 1980's with the introduction of Davis's Coxeter group construction (see Section 3.4 and [38]) and Gromov's definition and constructions of hyperbolic groups (Section 3.5).

For example, Theorem 3.7 implies that if π is a discrete group with a finite $K(\pi, 1)$ complex then π appears as a retractive subgroup in many Poincaré duality groups [131]. The Davis–Hausmann examples in Theorem 8.1 provided the first examples of Poincaré duality groups which cannot be realized by smooth closed aspherical manifolds.

Although these developments seem to raise the difficulty of the realization problem, as this is written few if any potential counterexamples are known for Problem 9.1.

The essential task in confirming that a construction yields a PD^n group is eased by recognition criteria for Poincaré duality groups, including this one [18, Theorem VIII.10.1]:

LEMMA 9.1. *Let Γ be a group of type FP. Γ is a* PD^n *group if and only if $H^i(\Gamma;\mathbb{Z}\Gamma)=0$ for all $i\neq n$ and $H^n(\Gamma;\mathbb{Z}\Gamma)=\mathbb{Z}$.*

The topological meaning of Lemma 9.1 is relatively clear if Γ is a group with a finite $K(\Gamma,1)$ complex $B\Gamma$, with universal covering space $E\Gamma$. In this case we have isomorphisms $H^i(\Gamma;\mathbb{Z}\Gamma)\cong H^i(B\Gamma;\mathbb{Z}\Gamma)\cong H^i_c(E\Gamma;\mathbb{Z})$, and the cohomology with compact supports of the contractible space $E\Gamma$ may be further identified with Čech cohomology of the end of $E\Gamma$. A topologist may thus wish to think of a PD^n group (with a finite classifying complex) as a group whose end is a Čech cohomology $(n-1)$-sphere: managing this computation during a uniformly described geometric construction is the key to Davis' examples (Section 3.4).

10. Group actions and stratified spaces

10.1. *Lie group actions*

Conner and Raymond showed that effective actions of compact Lie groups on closed aspherical manifolds are rather constrained [31, Theorem 5.6]:

THEOREM 10.1. *If (G,M) is a compact connected Lie group acting effectively on a closed aspherical manifold then*
(a) *G is a toral group with* $\dim(G)$ *at most the rank of the center of $\pi_1(M)$*;
(b) *all isotropy groups are finite*; *and*
(c) *the Euler characteristic $\chi(M)=0$.*

Several equivariant or fiberwise rigidity theorems for torus actions on aspherical manifolds are known, usually passing through structure theorems on these actions to conclude that if T^k acts smoothly and effectively on closed aspherical manifolds M^n and N^n and if M^n and N^n are homotopy equivalent, then these manifolds are equivariantly homeomorphic [31,32,123,124].

Another genre of uniqueness or rigidity theorem for group actions grows out of dynamics and Margulis super-rigidity. Zimmer and his students have worked with noncompact group actions and cocycle versions of superrigidity [187] while Katok and Spatzier have made progress on extending notions from the dynamics of Anosov flows to group actions [118].

10.2. *Finite groups, finite subgroups, and stratified rigidity*

Finite groups enter into the study of aspherical manifolds in several ways. The first is exemplified by an arithmetic group Γ acting on the contractible manifold G/K: although Γ has a finite-index subgroup which is torsion-free, arithmetic groups often contain elements of finite order, which act with fixed points. The quotient $\Gamma\backslash G/K$ is then an orbifold, rather than a manifold, and rigidity assertions about the action of Γ should engage with the stratified nature of this orbifold and the special role of finite subgroups.

The best studied class of groups from this point of view is the crystallographic groups. Yamasaki [186] did computations in the L-groups of crystallographic groups, following proposals of Quinn. Connolly and Kozniewski have worked on rigidity problems, obtaining both rigidity results [35] if the holonomy group is of odd order and some more technical hypotheses are satisfied. They have also found examples of the failure of equivariant rigidity [36] for crystallographic groups which have holonomy groups of even order.

Prassidis and Spieler [142] proved equivariant rigidity for every properly discontinuous, cocompact action (Γ, V) of a Coxeter group Γ on a contractible, boundaryless manifold V such that fixed point sets of finite subgroups are contractible and such that 3-dimensional fixed point sets are manifolds. This result was anticipated by the equivariant rigidity result of Rosas [153] in the right-angled case (all the exponents $m(i, j) = 2$ in the Coxeter presentation of Γ). (It is easier in the right-angled case than in general to describe the action of Γ on a universal Γ-space in terms of reflections across submanifolds. These groups also tend to split along square-root closed subgroups.) Prassidis and Spieler also obtain results on the equivariant Whitehead groups of such a manifold.

The formulation of rigidity conjectures is still in flux for a cocompact action of a discrete group Γ, containing torsion, on a contractible manifold V^n. Optimists have expressed a hope that if Γ contains no elements of even order and the fixed sets $\mathrm{Fix}(\Delta, V)$ are contractible (and perhaps the images of those fixed sets have well behaved fundamental groups, e.g., not $C_2 * C_2$), then (Γ, V) might be rigid. Pessimists point to the gap conditions of equivariant surgery as a sign of the distance remaining between known results and those hopes.

The Farrell–Jones isomorphism conjectures [78] are posed in terms of the classifying spaces for proper actions considered by Serre [156] and Connolly–Koźniewski [34]. Farrell and Jones enlarge the scope of these spaces by building a classifying space for actions with a prescribed class of isotropy subgroups (earlier work assumed finite isotropy, while Farrell and Jones emphasize virtually cyclic isotropy). The isomorphism conjectures are too complicated to state here and appear in [78, Section 1.7]: of particular interest is the attempt to formulate generalizations of the Borel or Novikov conjectures for stable pseudoisotopy theory, K-theory, and L-theory simultaneously.

A second context in which finite groups appear in the study of aspherical manifolds is in Nielsen-type problems: If M^n is a closed aspherical manifold and $G < \mathrm{Out}(\pi_1(M^n))$ is a finite group of outer automorphisms, does G lift to act on M^n? Conner and Raymond have studied these questions for some aspherical manifolds [32].

Finally, aspherical orbifolds may be regarded as stratified spaces, without special attention to the origins of the stratification in a group action. Much of the literature on stratified topology is actually concerned with equivariant problems, but foundational work on the topology of stratified spaces and related problems has been done by F. Quinn and B. Hughes, while stratified characteristic classes (especially in intersection homology) have been considered by R. MacPherson, S. Cappell, and J. Shaneson. The book of Goresky and MacPherson [95] serves as a guide to work on stratified spaces which derives from Thom and the algebraic geometers, in addition to making important contributions of its own. Equivariant surgery and related problems are a highly developed subject to which Petrie, Dovermann, Lück, Rothenberg, Schultz, and many others have contributed. Weinberger's notes on stratified spaces [182] are a snapshot of a rapidly developing subject.

11. Remarks on the literature

We begin with three general sources which have been especially useful in preparing this chapter. Two sets of lecture notes present the methods and viewpoint of Farrell, Hsiang, and Jones on the Borel Conjecture [73,50]. The mid-1990s understanding of the Novikov Conjecture is covered very well in the volumes [88,89]; the summary and historical survey [90] in the first volume may be recommended with particular enthusiasm.

A number of tools and problems in manifold topology are discussed in Ferry's book [87] and in the updated version of Kirby's problem list [120]. Ranicki's writings are the main sources on the algebraic theory of surgery and his algebraic reformulations of splitting obstructions should be consulted along with the original sources. Wall's book [175] remains the skeleton on which experts organize almost every discussion of non-simply connected surgery theory, but we still lack a satisfactory treatment of surgery for beginners. This is especially true for surgery in the topological category; see [122, pp. 264–289] for some of the details and see Ranicki's work, especially [149], for the current formulation of topological surgery. Freedman and Quinn should be consulted for four-dimensional surgery [92]. The two volumes of *Surveys on Surgery Theory* (Princeton University Press, 2000 and 2001) are a roadmap to the state of the art in surgery and its applications.

Apart from the work of M. Bökstedt, W.C. Hsiang, and I. Madsen (Theorem 5.1, [14]) on K-theory Novikov conjectures for groups with finitely generated homology in every dimension, the settled cases of the Borel Conjecture or Novikov Conjecture mix topology with geometrically or combinatorially defined classes of groups. See the chapters in this volume on geometric topics in group theory for more information on groups satisfying curvature conditions. Brown's book [18] is recommended for the homological and finiteness properties of groups mentioned above, and a forthcoming book of Geoghegan is probably the best reference on ends of groups.

ENR homology manifolds are discussed in [183] (Chapter 21 of this volume) and [19, 20]. These spaces, like much touched on in this chapter, remain the subject of vigorous investigation.

References

[1] J.F. Adams, *Infinite Loop Spaces*, Annals of Math. Studies, Vol. 90, Princeton University Press, Princeton (1978).

[2] D.R. Anderson, F.X. Connolly, S.C. Ferry and E.K. Pedersen, *Algebraic K-theory with continuous control at infinity*, J. Pure Appl. Alg. **94** (1994), 25–47.

[3] D.R. Anderson and H.J. Munkholm, *Boundedly Controlled Topology*, Lecture Notes in Math., Vol. 1323, Springer, Berlin (1988).

[4] D.V. Anosov, *Geodesic flows on closed Riemann manifolds with negative curvature*, Proc. Steklov Inst. Math. **90** (1960).

[5] O. Attie, J. Block and S. Weinberger, *Characteristic classes and distortion of diffeomorphisms*, J. Amer. Math. Soc. **5** (1992), 919–921.

[6] L. Auslander and F.E.A. Johnson, *On a conjecture of C.T.C. Wall*, J. London Math. Soc. **14** (1976), 331–332.

[7] W. Ballmann, M. Brin and P. Eberlein, *Structure of manifolds of nonpositive curvature. I*, Ann. of Math. **122** (1985), 171–203.

[8] W. Ballmann, M. Gromov and V. Schroeder, *Manifolds of Nonpositive Curvature*, Birkhäuser, Boston (1985).

[9] H. Bass, A. Heller and R. Swan, *The Whitehead group of a polynomial extension*, Inst. Hautes Études Sci. Publ. Math. **22** (1964), 64–79.

[10] D.J. Benson and P.H. Kropholler, *Cohomology of groups*, Handbook of Algebraic Topology, North-Holland, Amsterdam (1995), 917–950.

[11] T. Bedford, M. Keane and C. Series, *Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces*, Oxford University Press, Oxford (1991).

[12] M. Bestvina and N. Brady, *Morse theory and finiteness properties of groups*, Invent. Math. **129** (1997), 445–470.

[13] R. Bieri, *Gruppen mit Poincaré Dualität*, Comment. Math. Helv. **47** (1972), 373–396.

[14] M. Bökstedt, W.C. Hsiang and I. Madsen, *The cyclotomic trace and algebraic K-theory of spaces*, Invent. Math. **111** (1993), 465–540.

[15] A. Borel, *Compact Clifford–Klein forms of symmetric spaces*, Topology **2** (1963), 111–122.

[16] N. Bourbaki, *Groupes et Algèbres de Lie, Ch. 4–6*, Hermann, Paris (1968).

[17] W. Browder, *On the action of $\theta^n(\partial\pi)$*, Differential and Combinatorial Topology, Princeton University Press, Princeton (1965), 23–36.

[18] K.S. Brown, *Cohomology of Groups*, Graduate Texts in Math., Vol. 87, Springer, Berlin (1982).

[19] J.L. Bryant, S.C. Ferry, W. Mio and S. Weinberger, *Topology of homology manifolds*, Bull. Amer. Math. Soc. (N.S.) **28** (1993), 324–328.

[20] J.L. Bryant, S.C. Ferry, W. Mio and S. Weinberger, *Topology of homology manifolds*, Ann. of Math. **143** (1996), 435–467.

[21] S.E. Cappell, *Unitary nilpotent groups and Hermitian K-theory. I*, Bull. Amer. Math. Soc. **80** (1974), 1117–1122.

[22] S.E. Cappell, *Manifolds with fundamental group a generalized free product. I*, Bull. Amer. Math. Soc. **80** (1974), 1193–1198.

[23] S.E. Cappell, *A splitting theorem for manifolds*, Invent. Math. **33** (1976), 69–170.

[24] T.A. Chapman, *Topological invariance of Whitehead torsion*, Amer. J. Math. **93** (1974), 488–497.

[25] T.A. Chapman, *Controlled boundary and h-cobordism theorems*, Trans. Amer. Math. Soc. **280** (1983), 73–95.

[26] T.A. Chapman, *Controlled Simple Homotopy Theory and Applications*, Lecture Notes in Math., Vol. 1009, Springer, Berlin (1983).

[27] T.A. Chapman and S. Ferry, *Approximating homotopy equivalences by homeomorphisms*, Amer. J. Math. **101** (1979), 567–582.

[28] R.M. Charney and M.W. Davis, *Strict hyperbolization*, Topology **34** (1995), 329–350.

[29] M.M. Cohen, *A Course in Simple-Homotopy Theory*, Springer, New York (1973).

[30] E.H. Connell and J. Hollingsworth, *Geometric groups and Whitehead torsion*, Trans. Amer. Math. Soc. **140** (1969), 161–181.

[31] P.E. Conner and F. Raymond, *Actions of compact Lie groups on aspherical manifolds*, Topology of Manifolds (Athens, GA, 1969), J.C. Cantrell and C.H. Edwards, Jr., eds, Markham, Chicago (1970), 227–264.

[32] P.E. Conner and F. Raymond, *Deforming homotopy equivalences to homeomorphisms in aspherical manifolds*, Bull. Amer. Math. Soc. **83** (1977), 36–85.

[33] F.X. Connolly and M.O.M. Da Silva, *The groups $N^r K_0(\mathbb{Z}\pi)$ are finitely generated $\mathbb{Z}[\mathbb{N}^r]$ modules if π is a finite group*, K-Theory **9** (1995), 1–11.

[34] F. Connolly and T. Koźniewski, *Finiteness properties of classifying spaces of proper Γ-actions*, J. Pure Appl. Alg. **41** (1986), 17–36.

[35] F. Connolly and T. Koźniewski, *Rigidity and crystallographic groups, I*, Invent. Math. **99** (1990), 25–48.

[36] F. Connolly and T. Koźniewski, *Examples of lack of rigidity in crystallographic groups*, Algebraic Topology, Poznań 1989, S. Jackowski, B. Oliver and K. Pawalowski, eds, Lecture Notes in Math., Vol. 1474, Springer, Berlin (1991), 139–145.

[37] C.B. Croke, P. Eberlein and B. Kleiner, *Conjugacy and rigidity for nonpositively curved manifolds of higher rank*, Topology **35** (1996), 273–286.

[38] M.W. Davis, *Groups generated by reflections and aspherical manifolds not covered by Euclidean space*, Annals of Math. **117** (1983), 293–324.

[39] M.W. Davis, *Coxeter groups and aspherical manifolds*, Algebraic Topology, Aarhus 1982, I. Madsen, B. Oliver, eds, Lecture Notes in Math., Vol. 1051, Springer, Berlin (1984), 197–221.

[40] M. Davis and J.-C. Hausmann, *Aspherical manifolds without smooth or PL structures*, Algebraic Topology (Arcata, 1986), Lecture Notes in Math., Vol. 1370, Springer, New York (1989), 135–142.

[41] M.W. Davis and T. Januszkiewicz, *Hyperbolization of polyhedra*, J. Differential Geom. **34** (1991), 347–388.

[42] A.W.M. Dress, *Induction and structure theorems for orthogonal representations of finite groups*, Annals of Math. **102** (1975), 291–325.

[43] P. Eberlein and B. O'Neill, *Visibility manifolds*, Pacific J. Math. **46** (1973) 45–109.

[44] B. Eckmann and P. Linnell, *Poincaré duality groups of dimension two II*, Comment Math. Helv. **58** (1983), 111–114.

[45] B. Eckmann and H. Müller, *Poincaré duality groups of dimension two*, Comment Math. Helv. **55** (1980), 510–520.

[46] D. Epstein and M. Shub, *Expanding endomorphisms of flat manifolds*, Topology **7** (1968), 139–141.

[47] F.T. Farrell, *The obstruction to fibering a manifold over a circle*, Indiana Univ. Math. J. **21** (1971/1972), 315–346.

[48] F.T. Farrell, *The nonfiniteness of Nil*, Proc. Amer. Math. Soc. **65** (1977), 215–216.

[49] F.T. Farrell, *The exponent of UNil*, Topology **18** (1979), 305–312.

[50] F.T. Farrell, *Surgical Methods in Rigidity*, Tata Institute, Bombay (1996).

[51] F.T. Farrell and W.C. Hsiang, *Manifolds with* $\pi_1 = G \times_\alpha T$, Amer. J. Math. **95** (1973), 813–848.

[52] F.T. Farrell and W.C. Hsiang, *The topological-Euclidean space form problem*, Invent. Math. **45** (1978), 181–192.

[53] F.T. Farrell and W.C. Hsiang, *The Whitehead group of poly-(finite or cyclic) groups*, J. London Math. Soc. (2) **24** (1981), 308–324.

[54] F.T. Farrell and W.C. Hsiang, *On Novikov's conjecture for nonpositively curved manifolds. I*, Ann. of Math. **113** (1981), 199–209.

[55] F.T. Farrell and W.C. Hsiang, *The stable topological-hyperbolic space form problem for complete manifolds of finite volume*, Invent. Math. **69** (1982), 155–170.

[56] F.T. Farrell and W.C. Hsiang, *A structure set analogue of Chapman–Ferry–Quinn theory*, Continua, Decompositions, Manifolds (Austin, TX, 1980), Univ. Texas Press, Austin (1983), 172–190.

[57] F.T. Farrell and W.C. Hsiang, *Topological characterization of flat and almost flat Riemannian manifolds* M^n $(n \neq 3, 4)$, Amer. J. Math. **105** (1983), 641–672.

[58] F.T. Farrell and L.E. Jones, *Algebraic K-theory of hyperbolic manifolds*, Bull. Amer. Math. Soc. (N.S.) **14** (1986), 115–119.

[59] F.T. Farrell and L.E. Jones, *Algebraic K-theory of spaces stratified fibered over hyperbolic orbifolds*, Proc. Nat. Acad. Sci. USA **83** (1986), 5364–5366.

[60] F.T. Farrell and L.E. Jones, *h-cobordisms with foliated control*, Bull. Amer. Math. Soc. (N.S.) **15** (1986), 69–72. *Erratum*, Bull. Amer. Math. Soc. (N.S.) **16** (1987), 177.

[61] F.T. Farrell and L.E. Jones, *K-theory and dynamics. I*, Ann. of Math. **124** (1986), 531–569.

[62] F.T. Farrell and L.E. Jones, *K-theory and dynamics. II*, Ann. of Math. **126** (1987), 451–493.

[63] F.T. Farrell and L.E. Jones, *Algebraic K-theory of discrete subgroups of Lie groups*, Proc. Nat. Acad. Sci. USA **84** (1987), 3095–3096.

[64] F.T. Farrell and L.E. Jones, *Foliated control with hyperbolic leaves*, *K*-Theory **1** (1987), 337–359.

[65] F.T. Farrell and L.E. Jones, *Foliated control theory. I, II*, *K*-Theory **2** (1988), 357–430.

[66] F.T. Farrell and L.E. Jones, *The surgery L-groups of poly-(finite or cyclic) groups*, Invent. Math. **91** (1988), 559–586.

[67] F.T. Farrell and L.E. Jones, *Topological rigidity for hyperbolic manifolds*, Bull. Amer. Math. Soc. (N.S.) **19** (1988), 277–282.

[68] F.T. Farrell and L.E. Jones, *Implication of the geometrization conjecture for the algebraic K-theory of 3-manifolds*, Geometry and Topology (Athens, GA, 1985), C. McCrory and T. Shifrin, eds, Dekker, New York (1987), 109–113.

[69] F.T. Farrell and L.E. Jones, *A topological analogue of Mostow's rigidity theorem*, J. Amer. Math. Soc. **2** (1989), 257–370.

[70] F.T. Farrell and L.E. Jones, *Compact negatively curved manifolds (of dim $\neq 3, 4$) are topologically rigid*, Proc. Nat. Acad. Sci. USA **86** (1989), 3461–3463.
[71] F.T. Farrell and L.E. Jones, *Negatively curved manifolds with exotic smooth structures*, J. Amer. Math. Soc. **2** (1989), 899–908.
[72] F.T. Farrell and L.E. Jones, *Rigidity and other topological aspects of compact nonpositively curved manifolds*, Bull. Amer. Math. Soc. (N.S.) **22** (1990), 59–64.
[73] F.T. Farrell and L.E. Jones, *Classical Aspherical Manifolds*, CBMS Regional Conference Series in Mathematics, Vol. 75, Amer. Math. Soc., Providence, RI (1990).
[74] F.T. Farrell and L.E. Jones, *Foliated control without radius of injectivity restrictions*, Topology **30** (1991), 117–142.
[75] F.T. Farrell and L.E. Jones, *Computations of stable pseudoisotopy spaces for aspherical manifolds*, Algebraic Topology Poznań 1989, Lecture Notes in Math., Vol. 1474, Springer, Berlin (1991), 59–74.
[76] F.T. Farrell and L.E. Jones, *Rigidity in geometry and topology*, Proceedings of the International Congress of Mathematicians (Kyoto, 1990), Math. Soc. Japan, Tokyo (1991), 653–663.
[77] F.T. Farrell and L.E. Jones, *Stable pseudoisotopy spaces of compact nonpositively curved manifolds*, J. Differential Geom. **34** (1991), 769–834.
[78] F.T. Farrell and L.E. Jones, *Isomorphism conjectures in algebraic K-theory*, J. Amer. Math. Soc. **6** (1993), 249–297.
[79] F.T. Farrell and L.E. Jones, *Nonuniform hyperbolic lattices and exotic smooth structures*, J. Differential Geom. **38** (1993), 235–261.
[80] F.T. Farrell and L.E. Jones, *Topological rigidity for compact nonpositively curved manifolds*, Differential Geometry, Proc. Symp. Pure. Math., Vol. 54, Part 3, Amer. Math. Soc., Providence, RI (1993), 229–274.
[81] F.T. Farrell and L.E. Jones, *Complex hyperbolic manifolds and exotic smooth structures*, Invent. Math. **117** (1994), 57–74.
[82] F.T. Farrell and L.E. Jones, *The lower algebraic K-theory of virtually infinite cyclic groups*, K-Theory **9** (1995), 13–30.
[83] F.T. Farrell and L.E. Jones, *Some non-homeomorphic harmonic homotopy equivalences*, Bull. London Math. Soc. **28** (1996), 177–182.
[84] S. Ferry, *The homeomorphism group of a compact Hilbert cube manifold is an ANR*, Ann. of Math. **106** (1977), 101–119.
[85] S. Ferry, *Homotoping ϵ-maps to homeomorphisms*, Amer. J. Math. **101** (1979), 567–582.
[86] S. Ferry, *Controlled Topology and the Characterization of Topological Manifolds*, NSF-CBMS Conference Lecture Notes, Knoxville (1994).
[87] S.C. Ferry, *Topological Manifolds and Polyhedra*, Oxford University Press, to appear.
[88] S.C. Ferry, A. Ranicki and J. Rosenberg, eds, *Novikov Conjectures, Index Theorems and Rigidity, Vol. 1*, London Math. Soc. Lecture Notes, Vol. 226, Cambridge University Press, Cambridge (1995).
[89] S.C. Ferry, A. Ranicki and J. Rosenberg, eds, *Novikov Conjectures, Index Theorems and Rigidity, Vol. 2*, London Math. Soc. Lecture Notes, Vol. 227, Cambridge University Press, Cambridge (1995).
[90] S.C. Ferry, A. Ranicki and J. Rosenberg, *A history and survey of the Novikov conjecture*, Novikov Conjectures, Index Theorems and Rigidity, Vol. 1, S.C. Ferry, A. Ranicki and J. Rosenberg, eds, London Math. Soc. Lecture Notes, Vol. 226, Cambridge University Press, Cambridge (1995), 7–66.
[91] M.H. Freedman, *The topology of four-dimensional manifolds*, J. Differential Geom. **17** (1982), 357–453.
[92] M.H. Freedman and F. Quinn, *Topology of 4-Manifolds*, Princeton Mathematical Series, Vol. 39, Princeton University Press, Princeton (1990).
[93] M.H. Freedman and Z. Wang, *Controlled linear algebra*, Prospects in Topology, F. Quinn, ed., Annals of Math. Studies, Vol. 138, Princeton University Press, Princeton (1995), 138–156.
[94] K. Fukaya and T. Yamaguchi, *Almost nonpositively curved manifolds*, J. Differential Geom. **33** (1991), 67–90.
[95] M. Goresky and R. MacPherson, *Stratified Morse Theory*, Springer, Berlin (1988).
[96] J.A. Green, *Axiomatic representation theory for finite groups*, J. Pure Appl. Algebra **1** (1971), 41–77.
[97] M. Gromov, *Almost flat manifolds*, J. Differential Geom. **13** (1978), 231–241.
[98] M. Gromov, *Hyperbolic groups*, Essays in Group Theory, S.M. Gersten, ed., Math. Sci. Res. Inst. Publ., Vol. 8, Springer, New York (1987), 75–263.

[99] M. Gromov and I. Piatetski-Shapiro, *Non-arithmetic groups in Lobachevsky spaces*, Inst. Hautes Études Sci. Publ. Math. **66** (1988), 93–103.
[100] M. Gromov and W. Thurston, *Pinching constants for hyperbolic manifolds*, Invent. Math. **89** (1987), 1–12.
[101] J. Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. **84** (1986), 157–176.
[102] W.J. Harvey, *Boundary structure for the modular group*, Riemann Surfaces and Related Topics (Stony Brook, 1978), Ann. of Math. Studies, Vol. 97, Princeton University Press, Princeton (1978), 245–251.
[103] W.J. Harvey, *Geometric structure of surface mapping class groups*, Homological Group Theory, C.T.C. Wall, ed., London Math. Soc. Lecture Note Series, Vol. 36, Cambridge University Press, Cambridge (1979), 255–269.
[104] A. Hatcher, *Concordance spaces, higher simple homotopy theory, and applications*, Algebraic and Geometric Topology (Stanford, 1976), Proc. Symp. Pure Math., Vol. 32, Amer. Math. Soc., Providence, RI (1978), 2–21.
[105] A. Hatcher and J. Wagoner, *Pseudoisotopies of compact manifolds*, Astérisque **6** (1973).
[106] J. Hempel, 3-*Manifolds*, Annals of Mathematics Studies, Vol. 86, Princeton University Press, Princeton (1976).
[107] J.A. Hillman, *The Algebraic Characterization of* 4-*Manifolds*, London Math. Soc. Lecture Note Series, Vol. 198, Cambridge University Press, Cambridge (1994).
[108] W.C. Hsiang, *Borel's conjecture, Novikov's conjecture and the K-theoretic analogues*, Algebra, Analysis and Geometry (Taipei, 1988), World Scientific, Teaneck, NJ (1989), 39–58.
[109] W.C. Hsiang and J. Shaneson, *Fake tori*, Topology of Manifolds (Athens, GA, 1969), J.C. Cantrell and C.H. Edwards, Jr., eds, Markham, Chicago (1970), 18–51.
[110] W.-C. Hsiang and C.T.C. Wall, *On homotopy tori II*, Bull. London Math. Soc. **1** (1969), 341–342.
[111] B. Hu, *Whitehead groups of finite polyhedra with nonpositive curvature*, J. Differential Geom. **38** (1993), 501–517.
[112] B. Hu, *Relations between simplicial and manifold nonpositive curvature*, Geometric Group Theory (Proceedings of a Special Research Quarter at The Ohio State University, Spring 1992), R. Charney, M. Davis and M. Shapiro, eds, Ohio State University Math. Research Institute Publications, No. 3, de Gruyter, Berlin (1995).
[113] B. Hughes and A. Ranicki, *Ends of Complexes*, Cambridge Tracts in Mathematics, No. 123, Cambridge University Press, Cambridge (1996).
[114] J.E. Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge University Press (1990).
[115] F.E.A. Johnson, *On the realisability of polysurface groups*, J. Pure Appl. Algebra **15** (1979), 235–241.
[116] F.E.A. Johnson and C.T.C. Wall, *On groups satisfying Poincaré duality*, Ann. of Math. **96** (1972), 592–598.
[117] Y. Kamishima, K.B. Lee and F. Raymond, *The Seifert construction and its applications to infranilmanifolds*, Quart. J. Math. Oxford (2) **34** (1983), 433–452.
[118] A. Katok and R.J. Spatzier, *First cohomology of Anosov actions of higher rank Abelian groups and applications to rigidity*, Institut des Hautes Études Scientifiques. Publications Mathématiques/ **79** (1994), 131–156.
[119] M. Kervaire and J. Milnor, *Groups of homotopy spheres I*, Ann. of Math. **77** (1963), 504–537.
[120] R. Kirby, *Problems in low-dimensional topology*, Geometric Topology (Athens, GA, 1993), AMS/IP Studies in Advances Mathematics, No. 2, Part 2, Amer. Math. Soc./International Press, Providence, RI (1997), 35–473.
[121] R.C. Kirby and L.C. Siebenmann, *On the triangulation of manifolds and the Hauptvermutung*, Bull. Amer. Math. Soc. **75** (1969), 742–749.
[122] R.C. Kirby and L.C. Siebenmann, *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations*, Annals of Mathematics Studies, Vol. 88, Princeton University Press, Princeton (1977).
[123] K.B. Lee and F. Raymond, *Geometric realization of group extensions by the Seifert construction*, Combinatorics and Algebra, C. Greene, ed., Contemporary Math., Vol. 34, Amer. Math. Soc., Providence, RI (1984), 353–411.
[124] K.B. Lee and F. Raymond, *The role of Seifert fiber spaces in transformation groups*, Group Actions on Manifolds, R. Schultz, ed., Contemporary Math., Vol. 36, Amer. Math. Soc., Providence, RI (1985), 367–425.

[125] J.-L. Loday, *K-théorie algébrique et représentations de groupes*, Ann. Sci. École Norm. Sup. (4) **9** (1976), 309–377.

[126] A. Lubotzky and A.R. Magid, *Varieties of representations of finitely presented groups*, Mem. Amer. Math. Soc. **336** (1985).

[127] I. Madsen, C.B. Thomas and C.T.C. Wall, *The topological spherical form problem – II. Existence of free actions*, Topology **15** (1976), 375–382.

[128] A. Malcev, *On a class of homogeneous spaces*, Izv. Akad. Nauk SSSR Ser. Mat. **13** (1949), 9–32. Amer. Math. Soc. Transl. Series 1, Vol. 9 (1962), 277–307.

[129] G.A. Margulis, *Discrete groups of motions of manifolds of non-positive curvature*, Proc. Int. Congress. Math. (Vancouver, 1974), Amer. Math. Soc. Transl. **109** (1977), 33–45.

[130] G.A. Margulis, *Discrete Subgroups of Semisimple Lie Groups*, Springer, Berlin (1991).

[131] G. Mess, *Examples of Poincaré duality groups*, Proc. Amer. Math. Soc. **110** (1990), 1145–1146.

[132] J.W. Milnor, *On manifolds homeomorphic to the 7-sphere*, Ann. of Math. **64** (1956), 399–405.

[133] J.W. Milnor, *Hilbert's Problem 18: On crystallographic groups, fundamental domains, and on sphere packing*, Mathematical Developments Arising From Hilbert Problems, Proc. Symp. Pure Math. 28, Amer. Math. Soc., Providence, RI (1976), 491–506.

[134] G.D. Mostow, *Factor spaces of solvable groups*, Ann. of Math. **60** (1954), 1–27.

[135] G.D. Mostow, *Quasiconformal mappings in n-space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. **34** (1968), 53–104.

[136] G.D. Mostow, *Strong Rigidity of Locally Symmetric Spaces*, Annals of Math. Studies, Vol. 78, Princeton University Press, Princeton (1973).

[137] G. Moussong, *Hyperbolic Coxeter groups*, Ph.D. Thesis, Ohio State University (1988).

[138] A.J. Nicas, *Induction theorems for groups of homotopy manifold structures*, Mem. Amer. Math. Soc. **267** (1982).

[139] A.J. Nicas and C.W. Stark, *K-theory and surgery of codimension-two torus actions on aspherical manifolds*, J. London Math. Soc. (2) **31** (1985), 173–183.

[140] P. Orlik and H. Terao, *Arrangements of Hyperplanes*, Springer, Berlin (1992).

[141] E.K. Pedersen and C. Weibel, *A nonconnective delooping of algebraic K-theory*, Algebraic and Geometric Topology (Rutgers, 1983), A. Ranicki, N. Levitt and F. Quinn, eds, Lecture Notes in Math., Vol. 1126, Springer, Berlin (1985), 166–181.

[142] S. Prassidis and B. Spieler, *Rigidity of Coxeter Groups*, Preprint, Vanderbilt University (1996).

[143] F. Quinn, *Ends of maps*, I, Ann. of Math. **110** (1979), 275–331.

[144] F. Quinn, *Ends of maps*, II, Invent. Math. **68** (1982), 353–424.

[145] F. Quinn, *Ends of maps*. III: *Dimensions* 4 *and* 5, J. Differential Geom. **17** (1982), 503–521.

[146] F. Quinn, *Geometric algebra*, Algebraic and Geometric Topology (Rutgers, 1983), A. Ranicki, N. Levitt and F. Quinn, eds, Lecture Notes in Math., Vol. 1126, Springer, Berlin (1985), 182–198.

[147] M.S. Raghunathan, *Discrete Subgroups of Lie Groups*, Ergebnisse der Math., Vol. 68, Springer, Berlin (1972).

[148] A.A. Ranicki, *Exact Sequences in the Algebraic Theory of Surgery*, Princeton University Press, Princeton (1981).

[149] A.A. Ranicki, *Algebraic L-Theory and Topological Manifolds*, Cambridge Tracts in Mathematics, Vol. 102, Cambridge University Press, Cambridge (1992).

[150] A. Ranicki and M. Yamasaki, *Controlled K-Theory*, Topology Appl. **61** (1995), 1–59.

[151] F. Raymond and D. Wigner, *Construction of aspherical manifolds*, Geometric Applications of Homotopy Theory, Lecture Notes in Math., Vol. 657, Springer, Berlin (1978), 408–422.

[152] D. Rolfsen, *Knots and Links*, Publish or Perish, Wilmington, DE (1976).

[153] E. Rosas, *Rigidity theorems for right angled reflection groups*, Trans. Amer. Math. Soc. **308** (1988), 837–848.

[154] T.B. Rushing, *Topological Embeddings*, Pure and Applied Math., Vol. 52, Academic Press, London (1973).

[155] A. Selberg, *On discontinuous groups in higher-dimensional symmetric spaces*, Contributions to Function Theory (Int. Colloq. on Function Theory, Bombay, 1960), Tata Institute, Bombay (1960).

[156] J.-P. Serre, *Cohomologie des groupes discrets*, Prospects in Mathematics, Ann. of Math. Studies, Vol. 70, Princeton University Press, Princeton (1971), 77–169.

[157] J.-P. Serre, *Arithmetic groups*, Homological Group Theory, C.T.C. Wall, ed., London Math. Soc. Lecture Note Series, Vol. 36, Cambridge University Press, Cambridge (1979), 105–136.

[158] J.-P. Serre, *Trees*, Springer, Berlin (1980).

[159] J.L. Shaneson, *Wall's surgery obstruction groups for* $\mathbb{Z} \times G$, Ann. of Math. **90** (1969), 296–334.

[160] L. Siebenmann, *A total Whitehead obstruction to fibering over the circle*, Comment. Math. Helv. **45** (1970), 1–48.

[161] S. Smale, *Generalized Poincaré's conjecture in dimensions greater than four*, Ann. of Math. **74** (1961), 391–406.

[162] J. Stallings, *On fibering certain 3-Manifolds*, Topology of 3-Manifolds and Related Topics, M.K. Fort, Jr., ed., Prentice-Hall, Englewood Cliffs, NJ (1962), 95–100.

[163] J. Stallings, *The piecewise linear structure of Euclidean space*, Proc. Cambridge Philos. Soc. **58** (1962), 481–488.

[164] C.W. Stark, *L-theory and graphs of free Abelian groups*, J. Pure Appl. Algebra **47** (1987), 299–309.

[165] C.W. Stark, *A characterization of virtual Poincaré duality groups*, Michigan Math. J. **42** (1995), 99–102.

[166] C.B. Thomas and C.T.C. Wall, *The topological spherical space form problem I*, Compositio Math. **23** (1971), 101–114.

[167] V.G. Turaev, *Three-dimensional Poincaré complexes*, Math. USSR Sbornik **67** (1990), 261–282.

[168] E.B. Vinberg, *Hyperbolic reflection groups*, Russian Math. Surveys **40** (1985), 31–75.

[169] F. Waldhausen, *On irreducible 3-manifolds which are sufficiently large*, Ann. of Math. **87** (1968), 56–88.

[170] F. Waldhausen, *Algebraic K-theory of generalized free products*, Ann. of Math. **108** (1978), 135–256.

[171] F. Waldhausen, *Algebraic K-theory of topological spaces*, I, Algebraic and Geometric Topology (Stanford, 1976), Proc. Symp. Pure Math., Vol. 32, Amer. Math. Soc., Providence, RI (1978), 35–60.

[172] F. Waldhausen, *Algebraic K-theory of spaces, a manifold approach*, Current Trends in Algebraic Topology, Part 1 (London, Ontario, 1981), Canad. Math. Soc. Conf. Proc., Vol. 2, Amer. Math. Soc., Providence, RI (1982), 141–184.

[173] C.T.C. Wall, *Finiteness conditions for CW-complexes*, Ann. of Math. **81** (1965), 56–69.

[174] C.T.C. Wall, *On homotopy tori and the annulus theorem*, Bull. London Math. Soc. **1** (1969), 95–97.

[175] C.T.C. Wall, *Surgery on Compact Manifolds*, Academic Press, London (1970).

[176] C.T.C. Wall, *The topological space-form problems*, Topology of Manifolds (Athens, GA, 1969), J.C. Cantrell and C.H. Edwards, Jr., eds, Markham, Chicago (1970), 319–331.

[177] C.T.C. Wall, *List of problems*, Homological Group Theory, C.T.C. Wall, ed., LMS Lecture Notes, Vol. 36, Cambridge University Press, Cambridge (1979), 369–394.

[178] A. Weil, *Discrete subgroups of Lie groups* I, Ann. of Math. **72** (1960), 369–384.

[179] A. Weil, *Discrete subgroups of Lie groups* II, Ann. of Math. **75** (1962), 578–602.

[180] A. Weil, *Remarks on the cohomology of groups*, Ann. of Math. **80** (1964), 149–157.

[181] S. Weinberger, *On fibering four- and five-manifolds*, Israel J. Math. **59** (1987), 1–7.

[182] S. Weinberger, *The Topological Classification of Stratified Spaces*, University of Chicago Press, Chicago (1994).

[183] S. Weinberger, Homology manifolds, Handbook of Geometric Topology, R.J. Daverman and R.B. Sher, eds, Elsevier, Amsterdam (2002), 1085–1102.

[184] J.H.C. Whitehead, *A certain open manifold whose group is unity*, Quart. J. Math. **6** (1935), 268–279.

[185] J.A. Wolf, *Spaces of Constant Curvature*, (4th edn), Publish or Perish, Berkeley (1977).

[186] M. Yamasaki, *L-groups of crystallographic groups*, Invent. Math. **88** (1987), 571–602.

[187] R.J. Zimmer, *Ergodic Theory and Semisimple Groups*, Monographs in Math., Vol. 81, Birkhäuser, Boston (1984).

CHAPTER 21

Homology Manifolds

Shmuel Weinberger*

Department of Mathematics, University of Chicago, IL 60637, Chicago, USA

Contents

*Partially supported by an NSF grant.

HANDBOOK OF GEOMETRIC TOPOLOGY
Edited by R.J. Daverman and R.B. Sher

Generalized manifolds, for the purpose of this article, are finite-dimensional metric ANR homology manifolds, i.e., spaces X with $H_n(X, X - x) = \mathbb{Z}$ and $H_i(X, X - x) = 0$ for $i \neq n$, for all points $x \in X$. (See [5] for a weaker sounding equivalent condition.) These spaces arise naturally in a number of contexts, and, moreover, their study is intimately connected to fundamental problems in geometric topology and beyond.

There have been a number of expositions of the central ideas of this subject. I will just mention [19,20,11,9,30,44,52,58]. For the most part, I will try not to repeat the ideas discussed in these sources. I do this in an attempt to encourage readers to study these other sources whose authors (aside from that of the last cited reference) tend to have rather complementary perspectives.

1. Examples

Although the definition of a (topological) manifold is simple and natural, its very nature makes it hard to check directly: one is required to construct coordinate charts. A related, but seemingly different, complaint is that the admission fee for working in the topological category is quite steep: smooth and PL manifolds have pleasant tubular and regular neighborhood theories, Morse and handlebody theories, etc. In Top, it turns out that, after a great deal of effort, these tools do exist, except in dimension four – and, furthermore, in all dimensions, the structure of the theory is simpler than either Diff or PL. (See, e.g., [41,51,57].) Our thesis is that there is a natural class of spaces that are specified by possessing certain topological properties, and that for this class, while the admission price is yet steeper, the structure of the theory is even simpler.

But I am running ahead of myself. Let me first describe some places where homology manifolds arise naturally.

(A) Suppose one has a polyhedron and would like to decide if it is a manifold. What can one do? Since calculating homology and cohomology, etc. are algorithmic, one can check Poincaré duality. However, the polyhedron might accidentally satisfy PD, without being a manifold, like the letter T does. The next idea is to use Poincaré duality for open sets, and observe that the open set around the "join" point of the T does not satisfy duality. So, we are led to call a space a "generalized manifold" or "homology manifold" if every open subset satisfies proper Poincaré duality. A little sheaf theoretic argument shows that one only need look at open regular neighborhoods of simplices.

This condition is equivalent to the even easier to motivate requirement that for each $p \in X$, the local homology groups $H_*(X, X - p) \cong H_*(\mathbb{R}^n, \mathbb{R}^n - 0)$ are same as those for a manifold (e.g., those of Euclidean space).

Beyond this, it is hard to tell. For instance, given any homology sphere Σ, the open cone $c\Sigma$ (or its suspension $\Sigma\Sigma$!) is a homology manifold. It is easy to see that if Σ is nonsimply-connected, then this cannot be a manifold: the cone point has no simply-connected (small) deleted neighborhoods. If Σ is simply-connected, then $c\Sigma$ is a manifold: except for the cones of counterexamples to the 3-dimensional Poincaré conjecture, this follows from the truth of the Poincaré conjecture: such a Σ is a sphere. (The missing case follows from [36].)

On the other hand, there is no algorithm to determine if a homology n-sphere is simply-connected for n greater than four,[1] so one cannot algorithmically determine manifoldness among polyhedra. It would seem that is would only get worse when the singularities become nonisolated. In the next section we will see that that is not true. The singularities of a PL homology manifold P, from the topological point of view, are always isolated (theorem of Cannon and Edwards). However, it certainly reassured early topologists to know that all of their homotopical theorems remained true for homology manifolds.

Based on the observation of Kervaire that all PL homology spheres bound PL contractible manifolds (except in dimension 3, where Rochlin's theorem obstructs this), Cohen and Sullivan showed that the obstruction to finding a PL manifold V with a PL map $V \to P$ with contractible point inverses was an element of $H_{n-4}(P; \Theta_3)$,[2] where Θ_3 is the group of homology spheres up to homology h-cobordism. (Indeed, after one knows about Kirby–Siebenmann theory and the Cannon–Edwards theorem, it is not hard to use this analysis to figure out which topological manifolds are homeomorphic to polyhedra (as opposed to PL manifolds): see [38,43] for an approach not using the latter input.) The V produced will be homeomorphic to P iff P is a manifold.

(B) The topological version of such maps will play a central role in our discussion, and were studied by the great Bing school of geometric topologists over the course of several decades. Their point of view was typically opposite. Rather than starting with the homology manifold, one starts with a manifold V (usually it is Euclidean space) and then crushes various subsets to points to produce a new space, which we will not denote by P as it is almost never a polyhedron, but which we will denote as X.

Here is a well-known and important example. Consider a nullhomotopic embedding of a solid torus in itself as depicted in Figure 1.

We can re-embed a whole series of such solid tori in one another. Their intersection is an interesting subset, called *Wh*, after J.H.C. Whitehead who first considered its complement in the 3-sphere: it is nullhomotopic in any small neighborhood of itself, because the tori are nullhomotopic in one another. This implies that $\mathbb{R}^3/Wh$ is a homology manifold. It is certainly not a manifold, indeed the fundamental group of any small deleted neighborhood of *Wh* in $\mathbb{R}^3$ is infinitely generated and thus the same is true for the singular point of the constructed homology manifold!

Amazingly, according to Shapiro and Andrews–Rubin, the product $\mathbb{R} \times \mathbb{R}^3/Wh$ is $\mathbb{R}^4$! (The first published manifold factor is the more complicated example [4].) I should remark that the analysis of these "Bing doublings" of knots and associated decomposition spaces is at the heart of Freedman's proof of the four-dimensional Poincaré conjecture [36].

Other explicit decompositions gave rise to a space with no manifold points that is a manifold factor [53], or even one all of whose points have only nonsimply-connected deleted

[1] It is still unknown (but unlikely) whether recognition of the 4-sphere is possible. This is because one cannot produce arbitrary finitely-presented perfect groups with vanishing H_2 as the fundamental group of a homology 4-sphere [39]. The known general constructions require "balanced presentations" which are vanishingly rare among the outputs of machines for producing groups and presentations with logical complexity.

[2] We will always denote Borel–Moore, or locally finite, homology by the unadorned symbol H – "usual homology" will be viewed as homology with compact supports.

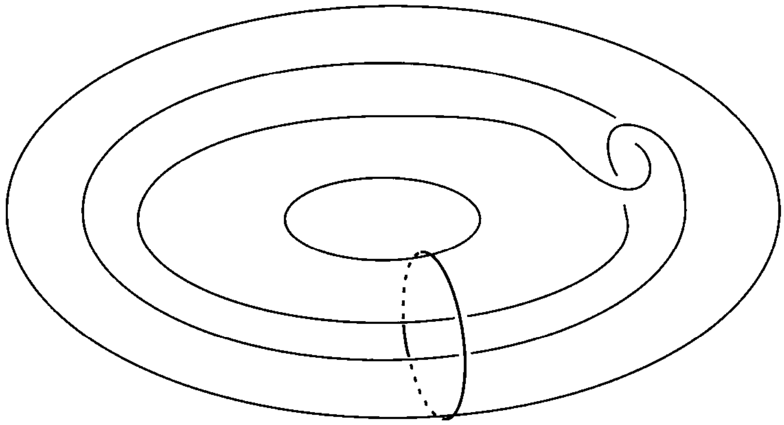


Fig. 1. First stage of the construction of the Whitehead continuum.

neighborhoods [13]. We shall discuss some of the other pathologies[3] below. I should also point out that in the early days of decomposition space theory, one produced, by carefully interlacing shrinkings around fixed-point sets, the first nonlinear group actions on Euclidean spaces (see [3,46]).

We note that this construction of homology manifolds produces them together with manifolds that very closely resemble them: We have a map $\varphi : M \to X$ which is a *hereditary homotopy equivalence*, i.e., we have that $\varphi|_{\varphi^{-1}(\mathcal{O})} : \varphi^{-1}(\mathcal{O}) \to \mathcal{O}$ is a (proper) homotopy equivalence for every open set $\mathcal{O} \subset X$. Such maps are called *CE maps*. M is said to *resolve* X, and φ is the *resolution*.

(C) A place where homology manifolds entered mathematics outside of pure topology was in the course of the proof of the following theorem:

THEOREM ([37]). *If $n > 3$, there are finitely many homeomorphism classes of Riemannian manifolds with any given upper bound on diameter and lower bounds on curvature and and volume.*

The point is that one shows that, if as one takes an infinite sequence of such manifolds, one can always extract a Gromov–Hausdorff convergent subsequence. A bit of work shows that in this case the limit is necessarily a homology manifold. Uniqueness of "approximate resolutions" (see Section 3) shows that all but finitely many manifolds in this subsequence are homeomorphic to each other, which proves the theorem.

For more information, see [47].

(D) Yet another place where homology manifolds arise naturally is in the theory of group actions on manifolds. I have already mentioned the examples of Bing and others of nonlinear actions on Euclidean spaces that stem from nonlocally flat fixed-point sets.

[3] Here pathology is not intended to be pejorative. I am speaking lovingly as would an immunologist. Indeed, the Bing school did develop many "splicing" techniques to create designer pathologies of various sorts. In studying their work, one can come to the point of view that the true subject here is the examples, not the theory built up surrounding them that I am focusing on in this survey.

In general, one can often deduce that a fixed-point set or the orbit space of the principal orbit type of a group action on a manifold is a homology manifold.[4] In my talk [58] I gave examples of results in fixed-point theory that would be obstructed if one only allowed manifold fixed-point sets, but become uniformly true when one allows (nonresolvable) homology manifolds as fixed sets.

A different sort of example is obtained by modding out $\mathbb{C}P^n$ by a decomposition of the sort discussed in (B), i.e., one which becomes "shrinkable" (the terminology will be explained in Section 2) after crossing with $\mathbb{R}$. Consider now, the circle bundle over this quotient space induced from the homotopy equivalence to $\mathbb{C}P^n$. The total space of this bundle is a manifold which the hypothesis on shrinkability identifies with S^{2n+1}. Rotating the fibers of this bundle gives a new circle action on that sphere. This action can easily be distinguished from the usual linear action on S^{2n+1} because its orbit space is not a manifold. On the other hand, the same shrinking fact easily yields the fact that for all finite subgroups $\mathbb{Z}_k \subset S^1$ the restricted actions are equivalent to the linear action. (To contrast, smoothly, one can show that a free action of S^1 on S^{2n+1} which is differentiably conjugate to a linear action for all of the finite subgroups, is in fact linear.)

2. Geometric methods

In this section, I would like to summarize some of the high points achieved by pure geometric reasoning, as distinguished from those to be discussed in Section 3 that require some algebraic organizational apparatus as well.

From the "classical period" I should probably mention the highlight theorems regarding dimension 2.

THEOREM (see [59]). *If X is a homology 2-manifold, then it is a 2-manifold.*

This was stated in a more quaint way in the old literature: e.g., a reasonable space which contains a circle and in which every circle separates the space into exactly 2 components is a 2-sphere (with a generalization to the surfaces possible, with care). Similarly, Montgomery had characterized 2-manifolds as homogeneous 2-dimensional ANR's during this classical period.

A result that seems less useful for the classification of manifolds, but is more in the spirit of later developments is the following result which paraphrases a result of R.L. Moore:

THEOREM. *Any CE map from a surface is a uniform limit of homeomorphisms.*

We say that uniform limits of homeomorphisms are *approximable by homeomorphisms*, and abbreviate this as ABH. (We leave it as an exercise to see that limits of homeomorphisms are CE.) Of course, we have seen above that this fails in dimension 3. Starting with Bing, a great deal of attention has been focused on the problem of seeing when a CE map is

[4] Actually, for some problems, one is also led by this route to consider mod p homology manifolds and other more exotic classes of spaces. For semifree circle actions the fixed-point set is a (possibly non-ANR) homology manifold.

ABH, and conversely, using funny decompositions to build designer homology manifolds with various properties. Almost all of this work rests on the fundamental:

BING SHRINKING CRITERION. A map $\varphi: X \to Y$ between compact metric spaces is ABH iff for every ε there is a self homeomorphism h of X which takes each point inverse of φ to a set of diameter $< \varepsilon$, and for which $d(\varphi, \varphi h) < \varepsilon$.

See [25,19] for an elegant (short) Baire category proof of this theorem. The decomposition of such an X into the preimages of φ is referred to as *shrinkable* when the requisite homeomorphisms h exist. Bing's criterion asserts that φ is ABH iff the decomposition is shrinkable.

It is quite instructive to convince oneself that the Whitehead decomposition is not shrinkable (although it is obvious that the quotient is not a manifold). The proof that after crossing with $\mathbb{R}$ the new decomposition is shrinkable is not at all hard (see [19, pp. 81–84]; it is all summarized in one picture on p. 83).

Besides its successes as a practical tool for deciding when various decompositions produce manifolds, one can wonder whether one is asking for too much in wanting the construction of the coordinate charts for our X to be close to the projection defining X. The following result of Siebenmann [54] shows that this is not at all the case:

THEOREM. *If M and X are manifolds of dimension > 4, then any CE $\varphi: M \to X$ is ABH.*

So shrinking the decomposition is equivalent to deciding manifoldness of the quotient. (Siebenmann's theorem is of significance for other reasons as well, as will be clear in the next section.) I should also mention that Armentrout had earlier proven the 3-dimensional version of this result assuming the Poincaré conjecture (indeed, it is equivalent to the Poincaré conjecture). The 4-dimensional case can be found in [49]. Siebenmann's motivation was the observation made by Sullivan that Novikov's proof of the topological invariance of rational Pontrjagin classes only used the hereditary homotopy equivalence property of homeomorphisms, not that they actually have unique point inverses.

The following result of Bob Edwards, conjectured by Jim Cannon following the verification of a number of important special cases (such as the double suspension problem), gives an extremely useful shrinkability criterion:

THEOREM ([25,19]). *If $\varphi: M \to X$ is a CE map, M a manifold of dimension > 4, and X is finite-dimensional (or ANR), then φ is ABH iff X has the DDP (disjoint disk property).*

DDP means that any two maps of D^2 into X can be arbitrarily closely approximated by maps whose images are disjoint from each other. It is quite easy to see that for, say, the cone on a nonsimply-connected homology sphere DDP fails: D^2's mapped in by radially extending nontrivial elements of π_1 cannot be moved disjoint from each other (indeed they cannot be moved disjoint from the cone point). On the other hand, once one crosses with $\mathbb{R}$, one can then slide in the $\mathbb{R}$ direction and make disks disjoint. Thus[5] one sees that:

[5] Modulo a little argument producing the resolution. However, aside from the 3-dimensional case, there is even a PL resolution, as we discussed in Section 1(A).

COROLLARY (Double suspension theorem). *The second suspension of any manifold homology sphere is a manifold (and hence the sphere).*

Thus, at least in high dimensions the problem of characterizing topological manifolds is reduced to the question of finding resolutions, to which we turn in the next section.

REMARK. I have given short shrift to a great deal of beautiful work that surrounds CE maps such as M. Brown's work on the Schoenflies theorem and taming theory for topological embeddings. For this, see, e.g., [20 and also 11,19].

3. Application of controlled topology

Controlled topology is the systematic reinvestigation of the classical problems of topology, such as the problem of putting boundaries on open manifolds (originally studied by Browder, Livesay, Levine, and Siebenmann), of making h-cobordisms into products (Smale, Barden, Mazur, Stallings), homotoping homotopy equivalences into homeomorphism (Kervaire, Milnor, Browder, Novikov, Sullivan, Wall), etc. with a view to keeping track of the size of the solutions of the problems.

Thus, one might want to know, for instance, how far a homotopy to a homeomorphism must move points during the course of the homotopy. Thought about this way, "controlling topology" should directly connect the type of geometric topology that has deep connections to algebraic K-theory and surgery (L-theory) to the type that studies whether maps are ABH, shrinkable, how embeddings might be tamed, etc., i.e., the type of question considered in the previous section. This is indeed the case.

Unfortunately for the novice there are many different types of control in the literature, each of which is best adapted to some problem or other: for instance, there are (epsilon) control, bounded control, and continuous control at infinity. Farrell and Jones [27] have considered "foliated control" to very good effect (see the chapter here on Topological Rigidity Theorems). We shall need "approximate" control in what follows, although, in some ways, it is the most difficult type of control of them all.

The most basic theorem in approximately controlled topology is the α-approximation theorem of Chapman and Ferry [16] (although this is an anachronistic description as it preceded the terminology by a good number of years[6]).

THEOREM. *For every open cover α of a manifold M there is a refinement β, such that* "β-controlled homotopy equivalences" *from a manifold N, $\varphi: N \to M$ are α-homotopic to homeomorphisms.*

A "β-controlled homotopy equivalence" φ, with homotopy inverse ψ, is one where the images in M of the paths followed by points during the course of the homotopies of $\varphi\psi$ and $\psi\varphi$ to the identities of M and N, respectively, always lie entirely in some element of

[6] It would be wrong for me not to mention here, though, the prescient paper [18] that did suggest some of the general context we now view as natural.

β. (Similarly for α-homotopic.) If M is compact, one can think of α and β as being small numbers measuring sizes.

Note that a CE map is a β-controlled homotopy equivalence for every β, so this theorem (essentially) includes Siebenmann's. Siebenmann's theorem is a "totally controlled" theorem, while the Chapman–Ferry theorem is the "approximately controlled" analogue. One approach to the α-approximation theorem is to obtain it from Siebenmann's theorem: show that for every α there is a β such that, for any γ, β-homotopy equivalences are α-homotopic to γ-homotopy equivalences. Letting γ get small very quickly enables one to check convergence of the sequence of maps to a continuous map, which is CE, and then by Siebenmann, ABH.

Moreover this latter approach, if done carefully, never uses the fact that the target is a manifold: one thus has a result to the effect that every "approximate resolution" is near a resolution; this refinement is necessary for the [37] result, discussed in Section 1.

Let us elucidate this point a bit. One of the important ideas in Quinn's landmark paper [48] was to liberate the control space from the problem being solved. That is one should introduce an auxiliary space P, just a finite-dimensional ANR, and require that our spaces M, N, X, whatever, all be equipped with maps to P. All of the covers, or distance measurements, will be made in P. (For the Chapman–Ferry theorem, the control map is $id: M \to M = P$.)

By allowing the control space to be an arbitrary homology manifold, one can consider the problem of homotoping a map that compares two different resolutions of the same X to be a (controlled) homeomorphism. After showing that controlled surgery theory exists, one can readily deduce from α-approximation that this is always possible (i.e., that a *uniqueness of resolution theorem* holds).[7]

Indeed, almost the same argument almost produces an existence of resolution theorem! (See [50,7].) The reason for this is that surgery theory works just as well to study the problem of when a Poincaré space is homotopy equivalent to a manifold as it does in deciding when a homotopy equivalence is homotopic to a homeomorphism (indeed the latter problem is studied as a special relative case of the former).

THEOREM. *Let X be a connected finite dimensional ANR homology manifold of dimension > 3, then there is an integer $i(X)$, which is locally defined (i.e., can be determined by restriction to any open subset) and which $= 1$ iff X can be resolved. $i(X)$ has several properties*:

(i) $i(e) = 1$,
(ii) $i(X) \equiv 1 \bmod 8$,
(iii) $i(X \times Y) = i(X)i(Y)$,
(iv) i *is invariant under CE maps.*

Indeed, it is often best to think of $i(X)$ as being the 0th Pontrjagin class of X. That is, recall that according to Novikov the rational Pontrjagin classes of manifolds are topological invariants. There are several possible approaches to the problem of directly defining

[7] Actually Quinn does not discuss controlled surgery in [48], but rather controlled h-cobordism. As a result, he requires more input (an h-cobordism rather than a controlled homotopy equivalence) to begin looking for homeomorphisms. This input he gets from Edwards' theorem.

Pontrjagin classes in a topologically invariant fashion: most traditionally, one can appeal to the theory of topological microbundles and use the fact that BSTop and BSO have the same rational homotopy to transport these classes from $H^*(BSO)$ to $H^*(BSTop)$ and then to the manifold using the classifying map of the topological tangent microbundle.

However, there is another approach, which is more closely related to what Novikov did. (Indeed, the pedigree goes back to Thom–Milnor and Rochlin before him who used such ideas in producing a PL invariant L-class for rational homology manifolds.) This method uses signatures and signatures of submanifolds to produce a homology class.[8] For manifolds one sees that this class in the top dimension is the fundamental class.

For general homology manifolds one has enough duality around on enough subsets to be able to apply the same procedure. The orientation class property is not so clear on inspection in general, and indeed it turns out not to be true. One could detect it though if one had a signature-type invariant of Poincaré spaces, which for manifolds involved the whole L-class, for then using the Poincaré structure of the homology manifold we would be able to see an invariant whose values would reflect the 0th (cohomology) L-class.

Finding such invariants of Poincaré spaces is exactly what the Novikov higher signature conjecture is about. (See the proceedings [33] for some surveys on this problem.) For instance, if $f : X \to T^n$ is a map, then the image of the L-class of X in $H_*(T^n)$ is a homotopy invariant according to old theorems of Farrell–Hsiang, Kasparov, and most important for us here Lusztig [42]. The advantage of Lusztig's approach is that he goes far enough to find an explicit homotopy invariant of Poincaré spaces with free abelian fundamental group, which for manifolds is the push-forward of the homology L-class (while the others merely show that that push-forward is a homotopy invariant).

Now, if M is a manifold of dimension n, and, say, the one-dimensional cohomology generates $H^n(M)$ under cup product, then one knows that the value of Lusztig's invariant is automatically the generator of $H_0(T^n)$. The signature of the transverse image of a point is necessarily one, because we have a degree one map to T^n and the point inverse image is a 0-manifold. What is at issue is whether this is necessarily the case for homology manifolds, and the answer is that it is not! There are nonresolvable homology manifolds not homotopy equivalent to any manifold. If one adds on a normalization axiom regarding the value of $i(M)$ for homology manifolds with free abelian fundamental groups, then this characterizes $i(X)$ (with the other axioms[9]).

Another definition of $i(X)$ for a homotopy sphere is simply this: if M is a manifold homotopy equivalent to $S^n \times \mathbb{C}P^2$ with $p_1 = 3k$, which is an approximate fibration over X,[10] then $i(X) = k$. In fact, for any X one can determine $i(X)$ from Pontrjagin classes of manifold approximate fibrations over X.

Rather than go through all of the details and repeat arguments that already exist in the literature, let me give a second description of where the $\mathbb{Z}$ obstruction comes from, and why it should be the only obstruction. (I'll engage in high-level philosophy, rather than even in heuristics.)

[8] More precisely, what is used is the theory of self-dual complexes of sheaves [14].

[9] The normalization and product axioms actually follow from a calculation for a torus and multiplicativity of Lusztig's invariant.

[10] This map is a generalized CE map: one wants the homotopy fiber of the restriction of the map to any open set to be independent of the open set – rather than be contractible, which is the CE condition.

A key property of surgery groups is their periodicity: $L_n(\square) \cong L_{n+4}(\square)$ whatever we put in the box. Now according to the α-approximation theorem there are no obstructions to making arbitrary controlled homotopy equivalences to M, controlled homotopic to homeomorphisms; in particular, there are no characteristic class obstructions. This means that the obstructions to getting the controlled homotopy equivalence must be rich enough to account for all the possible characteristic classes of all the maps that we might try to surger to a controlled equivalence. The aggregate of these is given by maps $[M : G/Top]$. By periodicity, we also see that the controlled surgery groups account for $[M \times D^4, \partial; G/Top]$ or D^{4k} for that matter. These sets of characteristic class sets are almost, but not quite, isomorphic: $\Omega^4(G/Top) \cong \mathbb{Z} \times G/Top$. The characteristic class theory is deficient in dimension 0; an integer which would be needed to have genuine periodicity – a periodicity which is forced by the structure of surgery theory to hold for the obstruction groups. This deficient $\mathbb{Z}$ is the $\mathbb{Z}$ that arises in the resolution theorem.

Also, since G/Top is a split summand of $\mathbb{Z} \times G/Top$ it is not surprising that one is not troubled in uniqueness theorems.

4. Toward a positive classification theory of homology manifolds

The previous section showed that many homology manifolds, namely those with $i(X) = 1$, are resolvable, and therefore one can recognize manifoldness for homology manifolds as being the same as DDP with $i(X) = 1$. It leaves open what one can say in general.

In [6,7] beginnings were made on this problem. In particular, the examples alluded to of homology manifolds not homotopy equivalent to resolvable ones were constructed. The main theorem of these papers can be phrased as:

THEOREM. *One can classify homology manifolds up to s-cobordism by means of surgery theory.*

This implies that for homology manifolds, one does have a fully periodic classification theory (above low dimensions). Let me state it carefully.

THEOREM. *If X is a homology manifold of dimension exceeding 5, then there is an isomorphism between the following two ("structure") sets: $S(X) \approx S(X \times D^4)$, where $S(X) = \{\varphi : Y \to X$, a simple homotopy equivalence, which is already a homeomorphism on any boundaries$\}/s$-cobordism and $S(X \times D^4)$ is the analogous set with $X \times D^4$ replacing X.*

Note the latter set is an abelian group (by exactly the same method one puts the group structure on homotopy groups: i.e., by "stacking")! Group structures were produced on topological manifold structure sets (see [55]), but periodicity fails as we discussed above.[11] Using homology manifold structure sets, one not only has a group structure, but much more: structure sets become covariant abelian group-valued functors: i.e., it is possible to

[11] Siebenmann observed that it was very close to being true.

push forward structures with respect to continuous maps between manifolds (or homology manifolds) whose dimensions differ by a multiple of four,[12] a process which one must admit is geometrically rather obscure.

Observe that if X is a manifold then the latter set is built up entirely out of manifolds (or at least resolvable homology manifolds, which are s-cobordant to manifolds by taking the mapping cylinder of a resolution map). Thus, homology manifolds fill in lacunae in the theory of manifolds. An important special case is:

$$S(S^n) \cong S(\mathbb{R}^n) \cong \mathbb{Z}.$$

The isomorphism is given by $(i(X) - 1)/8$. Thus, for instance, there are homotopy spheres other than the sphere in this category, but for each local index, it is unique up to s-cobordism, just as usual.

Obviously, one would want to know whether for local index other than 1 there are preferred local models which are, say, topologically homogeneous (for $i = 1$, these are the manifolds) and whether for these the s-cobordism theorem is true. I strongly believe this to be the case, but these problems have resisted a number of attempts.

To follow the outline suggested by Edwards' theorem, one might conjecture that the "good" models are determined by requiring the DDP. In [7] the construction produces DDP homology manifolds in every class of $S(X)$. In fact, in [8] we show that controlled surgery can be extended to DDP homology manifolds which implies, for instance:

COROLLARY. *Every homology manifold is resolvable by a DDP homology manifold.*

Let me briefly sketch the method by which these theorems are proven. All of the theorems are proven by (patching together suitable local versions of) the following argument: Suppose that we are given a Poincaré complex X which has vanishing total surgery obstruction, and we want to show it homotopy equivalent to a homology manifold. Then X has a normal invariant, i.e., there is a manifold M with a degree one normal map $M \to X$. Now, if the surgery obstruction of this vanished, then one could normally cobord this map to a homotopy equivalence from a manifold to X, and we would certainly be done. However, if the obstruction is nonzero, this is impossible, and we have to understand what it is that the condition of vanishing total surgery gives.

What it gives is this: the surgery obstruction must live in the image of $H_n(X; \mathbf{L}(e))$ under the assembly map. Now, this is the same thing as the controlled L-group $L^c(X \downarrow X)$, so we can act (in the sense of the action given in the surgery exact sequence, namely Wall realization) on the structures by this element on the structures of any manifold that UV^1 maps to X. The relevant manifold we choose is the boundary of a regular neighborhood of a 2-skeleton of M.

This using the new homotopy equivalence from the other side of the Wall realization to the boundary we can glue on this contraption between the sides of M separated by the boundary of this regular neighborhood. What we then obtain is a space mapping to

[12] This is somewhat reminiscent of the kind of conditions that Atiyah and Hirzebruch imposed to get wrong-way functoriality in K-theory: a small spin and/or complexity condition and the assumption that the manifolds have dimensions of the same parity.

X, whose surgery obstruction is now zero. Since we used controlled Wall realization, the space has good controlled Poincaré duality over X (although it certainly is not a homology manifold). A little thought shows that surgery theory can be applied to spaces that have the mild singularities of this one, so we obtain a controlled Poincaré complex, controlled over X, homotopy equivalent to X.

This is progress. What we would like to have is a space which is controlled Poincaré when measured over *itself*, i.e., a homology manifold, rather than just over our initial X. To do this we iterate the construction, with some care, so that each stage becomes controlled over the previous one, so that the limit is controlled over itself. This completes the sketch.

None of these methods produce homeomorphisms or even CE maps out of a given initially chosen homology manifold. Such would be necessary for proving homogeneity of these spaces or the s-cobordism theorem and the other standard tools of topological topology for DDP homology manifolds. In [6] we conjectured that all these and more are true. If correct, the picture of DDP homology manifolds will be very similar to that of ordinary topological manifolds.

REMARK. Some "evidence by analogy", based on experience from the theory of orbifolds in particular, and stratified spaces in general, for this collection of conjectures is suggested in [58]; this analogy extends the analogy between taming theory and decomposition spaces that [11] and [20] discuss. For instance, there are homogeneity, s-cobordism, local contractibility of homeomorphism group theorems for classes of spaces described purely in terms of local homotopical properties.

Personally, the only conjecture from the list in [6] that I have any doubts at all about is the local contractibility of homeomorphism group conjecture, and even that one I think is likely.

It is probably also worth mentioning that the analogies between surgery theory and index theory of elliptic operators that are developed in the appendix to chapter nine of [57], as well as in a number of papers in [33], work a lot better if one includes homology manifolds in surgery theory. (It works even better if one allows generalized orbifolds, where the strata are only homology manifolds, to mimic equivariant indices of operators.)

REMARK. A start on another basic geometric problem for homology manifolds, namely transversality, was made by Johnston in her thesis [40]. She shows, that while transversality to sub-homology manifolds of homology manifolds is obstructed, there are many cases where maps out of homology manifolds into manifolds are unobstructed. Moreover, her work shows that there are maps that are homotopic to transverse maps, but not approximable by such.

While these phenomena might make one wonder about how seriously to take the idea that "what difference should it make whether $i = 1$?", note, however, that DDP is what is necessary for general position, which is more used in building homeomorphisms than is transversality, per se, and that (again relying on our stratified experience) in the case of orbifolds [58], there are obstructions to transversality as well, but that ultimately, despite this, one can succeed in in proving all of the theorems conjectured to hold for homology manifolds.[13] However, it *does* show that there are interesting phenomena left to be discovered about their geometric nature where homology manifolds do not too closely parallel manifolds.

5. Remarks on infinite dimensions

A great deal of work has been done on infinite-dimensional manifolds of various sorts, some by methods intrinsic to that subject, others adaptations of the ideas of Section 2. The result most in the spirit of that section is:

TORUNCYK'S THEOREM. *An ANR is a Hilbert cube manifold iff it has the disjoint k-disk property for all k.*

To emphasize again one of our themes, here is a theorem that deduces homogeneity from local homotopical properties. The main reason that Edwards can get by with 2-disks and Toruncyk requires all disks is that homology manifolds have local Poincaré duality, while in infinite dimensions there is no corresponding principle.

More in the spirit of Section 3, one can study homology n-manifolds that have infinite covering dimension. That such spaces exist at all is itself a remarkable result of Dranishnikov [22], who had shown that finite cohomological dimension is consistent with infinite covering dimension. His examples are cell-like images of finite-dimensional spaces (a theorem of Edwards forces this). He detects their infinite dimensionality by showing that they and their finite-dimensional "resolutions" have different K-theory.

(It is important here to use a nonconnective theory like K-theory because for ordinary homology and cohomology, the Vietoris–Begle theorem gives an equivalence, and one can deduce the case of any connective theory from this.)

Not so much is known about infinite-dimensional homology n-manifolds. It is quite easy to see that they can possess a dearth of finite-dimensional subspaces, and have other counterintuitive features. On the other hand, they do arise fairly naturally in some contexts: A theorem of Engel–Moore produces such examples as Gromov–Hausdorff limits of manifolds with some uniform contractibility function (class $\mathrm{LC}(\rho)$).

Subsequently, Ferry [29] established a simple homotopy theory for such limits, and showed that every compact homology n-manifold has at most finitely many resolutions [28]. This can be viewed as saying that only finitely many $\mathrm{LC}(\rho)$ manifolds can mutually degenerate in Gromov–Hausdorff space. Such theorems can be viewed as natural extensions of the theorem of [37] discussed above.

A really new wrinkle then arises: there are examples where there are distinct resolutions (always an odd number, because G/Top is Eilenberg–MacLane at 2, and is therefore essentially connective as far as convergence properties of spectral sequences are concerned). These were found by Dranishnikov and Ferry [23]. The same method also produces homology manifolds without finite-dimensional resolutions (by closed homology manifolds). The basic point behind the construction is that controlled surgery is a homology theory

[13] It did, however, interfere with with earlier attempts at classification results (notably the tour de force by [45]) which only succeeded for odd-order groups where transversality is shown to be possible.

closely related at odd primes to periodic K-theory. The difference between the Htheories of X and its resolving manifold gives a place for different resolutions to come from.

Finally, in [24], these methods are applied to produce interesting noncompact finite-dimensional Riemannian manifolds which contradict some of the standard "large scale" conjectures related to the Novikov conjecture. For instance, a uniformly contractible manifold is constructed there, which is not boundedly topologically rigid. (Universal covers of aspherical manifolds are always uniformly contractible, so the rigidity of such manifolds is an analogue of the Borel rigidity conjecture for aspherical manifolds – and its verification for many manifolds, such as Hadamard manifolds, is not that difficult and is an important tool in proving the Novikov conjecture for the compact aspherical manifolds they cover; see [35].) The homology manifold that arises in this example is a Gromov–Hausdorff limit of rescaled versions of the spheres of various radii and is a cell-like image of the sphere whose K-theory differs from that of the usual sphere.

While it is not yet clear whether the point-set geometric topology of these spaces has any features to recommend it, it seems that these spaces do arise from time to time, and do deserve study, at least to the extent of using them as control spaces, and maybe even to the point of studying them under some less restrictive equivalence relation that homeomorphism.

6. Some problems

This is not, by any means, intended to be a comprehensive list of problems. It is a list of questions some of which I think are actually important, and some others of which one can make progress on.

(A) Homogeneity (etc.). This was mention in Section 4. The problem is to show that there are nice homogeneous models for k-dimensional homology manifolds with given local index $= i$. "Manifolds" modelled on these should satisfy the s-cobordism theorem.

Presumably, the locally best examples will be characterized by satisfying the DDP, although it is conceivable that some other property will pick out the "best examples" and only later will it end up being equivalent to DDP.

In other words, one might find the good models somewhere, without being able to prove right away the analogue of the double suspension theorem or Edwards' theorem. Probably it would be a mistake to think of that as a failure.

(B) $\times\mathbb{R}$ problem. One does not know whether or not every decomposition of a manifold becomes shrinkable after crossing with a single copy of $\mathbb{R}$. Daverman has long ago shown that 2 copies are enough. Even a moron can show that 5 are enough.

(C) Systematic classification of nonlocally flat embeddings. Much of the theory discussed in Section 3 grew out of the problems of producing mapping cylinder structures for submanifolds, or taming them up to approximation. I think it might be interesting to study nonlocally flat embeddings with some given measure of nonlocal flatness, say in terms of a perfect group which local fundamental groups are given a map to. (This is a measure of nonlocal flatness because one knows that local flatness in codimension not two is the same as local simple connectivity.) For instance, it is not hard to see that while all embeddings of S^n in S^{n+c}, $c > 2$, can be approximated by locally flat ones (which are of course unique up to isotopy), they are not all isotopic if we require that the complement always have a map to, say, the dodecahedral group. (The circles occurring from double suspensions of different homology spheres are often not concordant in this sense.)

I am not sure whether every concordance class has a homogeneous representative (this could be especially interesting in codimension 3).

(D) Low dimensions. I have mentioned that the theory in dimension three has connections to the Poincaré conjecture. See [52] for more survey of dimension three. In dimension four, one does not even have a conjectured characterization of manifolds. This is clearly a matter of prime importance.

(E) Infinite dimensions. It would be good to have an understanding of which infinite dimensional homology n-manifolds can be resolved, and by how many manifolds. The point-set theory here seems to be wide open.

(F) Mod p-homology manifolds or $\mathbb{Q}$-homology manifolds. Here very little is known. If one starts by thinking about the polyhedral case, one sees immediate difficulties because the group of homology n-spheres up to homology h-cobordism is rather large and complicated (see [1] and [2]), unlike the case for integral homology spheres.

However, in transformation groups these spaces arise naturally, and it would be good to have at least obstructions to resolution for them and surgery up to s-cobordism.

Incidentally, for ANR $\mathbb{Q}$-homology manifolds, one can still define the local index. I very much doubt that $i(X) \equiv 1 \bmod 8$ still holds, but I haven't succeeded in producing examples.

References

[1] J.P. Alexander, G.C. Hamrick and J.W. Vick, *Involutions on homotopy spheres*, Invent. Math. **24** (1974), 35–50.

[2] J. Barge, J. Lannes, F. Latour and P. Vogel, *Λ-sphères*, Ann. Sci. École Norm. Sup. (4) **7** (1974), 463–505 (French).

[3] R.H. Bing, *A homeomorphism between the* 3*-sphere and the sum of two solid horned spheres*, Ann. of Math. (2) **56** (1952), 354–362.

[4] R.H. Bing, *The cartesian product of a certain nonmanifold and a line is* E^4, Ann. of Math. (2) **70** (1959), 399–412.

[5] G.E. Bredon, *Generalized manifolds,* Topology of Manifolds (Athens, GA, 1969), Markham, Chicago, IL (1970), 461–469.

[6] J. Bryant, S. Ferry, W. Mio and S. Weinberger, *Topology of homology manifolds*, Bull. Amer. Math. Soc. (N.S.) **28** (2) (1993), 324–328.

[7] J. Bryant, S. Ferry, W. Mio and S. Weinberger, *Topology of homology manifolds*, Ann. of Math. (2) **143** (3) (1996), 435–467.

[8] J. Bryant, S. Ferry, W. Mio and S. Weinberger, *Disjoint disk resolutions of homology manifolds*, In preparation.

[9] J.W. Cannon, *The recognition problem: what is a topological manifold*, Bull. Amer. Math. Soc. **841** (1978), 57–102.

[10] J.W. Cannon, *Shrinking cell-like decompositions of manifolds: Codimension three*, Ann. of Math. (2) **110** (1) (1979), 83–112.

[11] J.W. Cannon, *The characterization of topological manifolds of dimension* $n < 5$, Proceedings of the International Congress of Mathematicians (Helsinki, 1978), Acad. Sci. Fennica, Helsinki (1980), 449–454.

[12] J.W. Cannon, J.L. Bryant and R.C. Lacher, *The structure of generalized manifolds having nonmanifold set of trivial dimension*, Geometric Topology (Proc. Georgia Topology Conf., Athens, GA, 1977), Academic Press, New York (1979), 261–300.

[13] J.W. Cannon and R.J. Daverman, *A totally wild flow*, Indiana Univ. Math. J. **30** (3) (1981), 371–387.
[14] S. Cappell, J. Shaneson and S. Weinberger, *Classes topologiques caracteristiques pour les actions de groupes sur les espaces singuliers*, C. R. Acad. Sci. Paris Sér. I Math. **313** (5) (1991), 293–295.
[15] T.A. Chapman, *Approximation results in topological manifolds*, Mem. Amer. Math. Soc. **34** (251) (1981).
[16] T.A. Chapman and S. Ferry, *Approximating homotopy equivalences by homeomorphisms*, Amer. J. Math. **101** (3) (1979), 583–607.
[17] M. Cohen, *Simplicial structures and transverse cellularity*, Ann. of Math. (2) **85** (1967), 218–245.
[18] E. Connell and J. Hollingsworth, *Geometric groups and Whitehead torsion*, Trans. Amer. Math. Soc. **140** (1969), 161–181.
[19] R. Daverman, *Decompositions of Manifolds*, Pure and Applied Mathematics, Vol. 124, Academic Press, Orlando, FL (1986).
[20] R. Daverman, *The intimate connections among decomposition theory, embedding theory, and manifold structure theory*, Geometric Topology and Shape Theory (Dubrovnik, 1986), Lecture Notes in Math., Vol. 1283, Springer, Berlin (1987), 43–47.
[21] R. Daverman and J.J. Walsh, *A ghastly generalized n-manifold*, Illinois J. Math. **25** (4) (1981), 555–576.
[22] A.N. Dranishnikov, *K-theory of Eilenberg–MacLane spaces and cell-like mapping problem*, Trans. Amer. Math. Soc. **335** (1) (1993), 91–103.
[23] A.N. Dranishnikov and S. Ferry, *Convergence of $LC(\rho)$ manifolds in Gromov–Hausdorff space*, Preprint.
[24] A.N. Dranishnikov and S. Weinberger, *A flexible uniformly contractible manifold*, Preprint.
[25] R.D. Edwards, *The topology of manifolds and cell-like maps*, Proceedings of the International Congress of Mathematicians (Helsinki, 1978), Acad. Sci. Fennica, Helsinki (1980), 111–127.
[26] R.D. Edwards, *Characterizing infinite-dimensional manifolds topologically* (after Henryk Torunczyk), Seminaire Bourbaki (1978/79), Exp. no. 540, Lecture Notes in Math., Vol. 770, Springer, Berlin (1980), 278–302.
[27] F.T. Farrell and L. Jones, *Rigidity in geometry and topology*, Proceedings of the International Congress of Mathematicians, Vol. I (Kyoto, 1990), Math. Soc. Japan, Tokyo (1991), 653–663.
[28] S. Ferry, *Topological finiteness theorems in Gromov–Hausdorff space*, Duke Math. J. **74** (1994), 95–106.
[29] S. Ferry, *Limits of polyhedra in Gromov–Hausdorff space*, Topology **37** (1998), 1325–1338.
[30] S. Ferry, *CBMS lecture notes on homology manifolds*, In preparation.
[31] S. Ferry, *Notes on Top Topology*, Oxford University Press, to appear.
[32] S. Ferry and E.K. Pedersen, *Epsilon surgery theory*, Novikov Conjectures, Index Theorems and Rigidity, Vol. 2 (Oberwolfach, 1993), London Math. Soc. Lecture Note Ser., Vol. 227, Cambridge Univ. Press, Cambridge (1995), 167–226.
[33] S. Ferry, A. Ranicki, and J. Rosenberg, eds, *Novikov Conjectures, Index Theorems and Rigidity*, 2 Vol. (Oberwolfach, 1993), London Math. Soc. Lecture Note Ser., Vols 226–227, Cambridge Univ. Press, Cambridge (1995).
[34] S. Ferry and S.Weinberger, *Curvature, tangentiality, and controlled topology*, Invent. Math. **105** (2) (1991), 401–414.
[35] S. Ferry and S.Weinberger, *A coarse approach to the Novikov conjecture, Novikov Conjectures*, Index Theorems and Rigidity, Vol. 1 (Oberwolfach, 1993), London Math. Soc. Lecture Note Ser., Vol. 226, Cambridge Univ. Press, Cambridge (1995), 147–163.
[36] M. Freedman, *The topology of four-dimensional manifolds*, J. Differential Geom. **17** (3) (1982), 357–453.
[37] K. Grove, P. Petersen and J. Wu, *Geometric finiteness theorems via controlled topology*, Invent. Math. **99** (1) (1990), 205–213.
[38] R. Galewski and R. Stern, *Classification of simplicial triangulations of topological manifolds*, Ann. of Math. (2) **111** (1) (1980), 1–34.
[39] J.Cl. Hausmann and S. Weinberger, *Caracteristiques d'Euler et groupes fondamentaux des varietés de dimension* 4, Comment. Math. Helv. **60** (1) (1985), 139–144.
[40] H. Johnston, *Transversality for some non-manifold settings*, Ph.D. thesis, University of Chicago (1997).
[41] R.C. Kirby, L.C. Siebenmann and L.C. Laurence, *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations* (with notes by J. Milnor and M. Atiyah, Annals of Mathematics Studies, Vol. 88), Princeton University Press, Princeton, NJ; University of Tokyo Press, Tokyo (1977).
[42] G. Lusztig, *Novikov's higher signature and families of elliptic operators*, J. Differential Geom. **7** (1972), 229–256.

[43] T. Matumoto, *Triangulation of manifolds*, Algebraic and Geometric Topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, CA, 1976), Part 2, Proc. Sympos. Pure Math., Vol. 22, Amer. Math. Soc., Providence, RI (1978), 3–6.
[44] W. Mio, *Homology Manifolds*, The Wall Festschrift, Princeton University Press, to appear.
[45] I. Madsen and M. Rothenberg, *On the classification of G-spheres*, Proceedings of the Northwestern Homotopy Theory Conference (Evanston, IL, 1982), Comtemp. Math., Vol. 19, Amer. Math. Soc, Providence, RI (1983), 193–226.
[46] D. Montgomery and L. Zippin, *Topological Transformation Groups*, Robert E. Krieger Publishing Co., Huntington, NY (1974) (reprint of the 1955 original).
[47] P. Petersen, *Gromov–Hausdorff convergence of metric spaces*, Differential Geometry: Riemannian Geometry (Los Angeles, CA, 1990), Proc. Sympos. Pure Math., Vol. 54, Part 3, Amer. Math. Soc., Providence, RI (1993), 489–504.
[48] F. Quinn, *Ends of maps*, I, Ann. of Math. (2) **110** (2) (1979), 275–331.
[49] F. Quinn, *Ends of maps*, III. *Dimensions* 4 *and* 5, J. Differential Geom. **17** (3) (1982), 503–521.
[50] F. Quinn, *Resolutions of homology manifolds, and the topological characterization of manifolds*, Invent. Math. **72** (2) (1983), 267–284.
[51] A. Ranicki, *The total surgery obstruction*, Algebraic Topology, Aarhus 1978 (Proc. Sympos. Univ. Aarhus, Aarhus, 1978), Lecture Notes in Math., Vol. 763, Springer, Berlin (1979), 275–316.
[52] D. Repovš, *The recognition problem for topological manifolds: A survey*, Workshop on Geometry and Topology (Hanoi, 1993), Kodai Math. J. **17** (3) (1994), 538–548.
[53] R. Rosen, *E^4 is the Cartesian product of a totally non-Euclidean space and E^1*, Ann. of Math. (2) **73** (1961), 349–361.
[54] L. Siebenmann, *Approximating cellular maps by homeomorphisms*, Topology **11** (1972), 271–294.
[55] L. Siebenmann, *The Periodicity in Topological Surgery*, Foundational Essays on Topological Manifolds, Smoothings, and Triangulations, University of Tokyo Press, Tokyo (1977).
[56] C.T.C. Wall, *Surgery on Compact Manifolds*, London Mathematical Society Monographs, Vol. 1, Academic Press, London (1970).
[57] S. Weinberger, *The Topological Classification of Stratified Spaces*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL (1994).
[58] S. Weinberger, *Nonlocally linear manifolds and orbifolds*, Proceedings of the International Congress of Mathematicians, Vols 1, 2 (Zurich, 1994), Birkhäuser, Basel (1995), 637–647.
[59] R. Wilder, *Topology of Manifolds*, Topology of Manifolds, American Mathematical Society Colloquium Publications, Vol. 32, Amer. Math. Soc., Providence, RI (1979) (Reprint of 1963 edition).

Author Index

Roman numbers refer to pages on which the author (or his/her work) is mentioned. Italic numbers refer to reference pages. Numbers between brackets are the reference numbers. No distinction is made between first and co-author(s).

Subject Index