Lecture 5: Instanton moduli space

The steady progress of physics requires for its theoretical formulation a mathematics that gets continually more advanced.

— PAM Dirac, 1931

In the previous lecture we constructed a $k = 1$ SU(2) instanton on S^4 and in fact saw that it belongs to a five-parameter family of such instantons. This is not an accident and in this lecture we will see that there is a moduli space of instantons, which is a disjoint union of a countable number of finite-dimensional connected subspaces labelled by the instanton number. To a first approximation, the moduli space is the quotient of the space of (anti)self-dual connection modulo gauge transformations. However this space turns out to be singular in general and in order to guarantee a smooth quotient we will have either to restrict ourselves to irreducible connections, or else quotient by a (cofinite) subgroup of gauge transformations.

5.1 Irreducible connections

Throughout this section we will let $P \to M$ be a fixed principal G-bundle with connection $H \subset TP$. Let ω denote the connection 1-form. A smooth curve $\tilde{\gamma} : [0,1] \to P$ is said to be **horizontal** if the velocity vector is everywhere horizontal: $\dot{\tilde{\gamma}}(t) \in H_{\tilde{\gamma}(t)}$ for all t. This is equivalent to $\omega(\dot{\tilde{\gamma}}(t)) = 0$. Let $\gamma(t) = \pi(\tilde{\gamma}(t))$ denote the projection of the curve onto M. Assume that the curve is small enough so that the image of γ lies inside some trivialising neighbourhood U_α. Then $\psi_\alpha(\tilde{\gamma}(t)) = (\gamma(t), g(t))$, where $g(t)$ is a smooth curve on G.

Done? ❑ **Exercise 5.1.** Show that the condition $\omega(\dot{\tilde{\gamma}}(t)) = 0$ translates into the following ordinary differential equation for the curve $g(t)$:

$$\mathrm{ad}_{g(t)^{-1}} A_\alpha(\dot{\gamma}(t)) + (g^*\theta)(\dot{\gamma}(t)) = 0 \,, \tag{20}$$

where A_α is the gauge field on U_α corresponding to the connection, and θ is the left-invariant Maurer–Cartan 1-form on G. Show further that for a matrix group, this equation becomes

$$\dot{g}(t) + A_\alpha(\dot{\gamma}(t))g(t) = 0 \,. \tag{21}$$

Being a first-order ordinary differential equations with smooth coefficients, equation (20) (equivalently (21)) has a unique solution for specified initial conditions, so that if we specify $g(0)$ then $g(1)$ is determined uniquely. This then defines a map $\Pi_\gamma : P_{\gamma(0)} \to P_{\gamma(1)}$ from the fibre over $\gamma(0)$ to the fibre over $\gamma(1)$, associated to the curve $\gamma : [0,1] \to M$. Rephrasing, given the curve γ, there is a unique horizontal lift $\tilde{\gamma}$ once we specify $\tilde{\gamma}(0) \in P_{\gamma(0)}$ and $\Pi_\gamma \tilde{\gamma}(0) = \tilde{\gamma}(1)$ is simply the endpoint of this horizontal curve. The map Π_γ is called **parallel transport along** γ **with respect to the connection** H.

Now let γ be a loop, so that $\gamma(0) = \gamma(1)$. Parallel transport along γ defines a group element $g_\gamma \in G$ defined by $g_\gamma = g(1)g(0)^{-1}$. To show that this element is well-defined, we need to show that it does not depend on the initial point $g(0)$. Indeed, suppose we choose a different starting point $\overline{g}(0)$. Then there is some group element $h \in G$ such that $\overline{g}(0) = g(0)h$. The parallel-transport equations (20) and (21) are clearly invariant under the right G action, whence $\overline{g}(t) := g(t)h$ solves the equation with initial condition $\overline{g}(0)$. Therefore the final point of the curve is $\overline{g}(1) = g(1)h$, whence $\overline{g}(1)\overline{g}(0)^{-1} = g(1)g(0)^{-1}$ and g_γ is well-defined. This defines a map from piecewise-smooth loops based at $m = \gamma(0)$ to G, whose image is a subgroup of G called the **holonomy group of the connection at** m denoted

$$\mathrm{Hol}_m(\omega) = \left\{ g_\gamma \middle| \gamma : [0,1] \to M \,,\; \gamma(1) = \gamma(0) = m \right\} \,. \tag{22}$$

Done? ❑ **Exercise 5.2.** Show that the holonomy group is indeed a subgroup of G; that is, show that it is closed under group multiplication. More precisely, if g_{γ_1} and g_{γ_2} are elements in $\mathrm{Hol}_m(\omega)$, then show that so is their product by exhibiting a loop γ based at m such that $g_\gamma = g_{\gamma_1} g_{\gamma_2}$. Further show that if $m, m' \in M$

belong to the same connected component, the holonomy groups $\mathrm{Hol}_m(\omega)$ and $\mathrm{Hol}_{m'}(\omega)$ are conjugate in G and hence isomorphic.

It follows from the previous exercise, that if M is connected, then the holonomy group of the connection is well-defined as a conjugacy class of subgroups of G. A connection is said to be **irreducible** if the holonomy group is precisely G and not a proper subgroup. The importance of the concept of irreducibility is that the group $\mathscr{G}$ of gauge transformations acts (almost) freely on the space of irreducible connections. The key observation is the covariance of parallel transport under gauge transformations.

Done? ❑ **Exercise 5.3.** Let $\Phi \in \mathscr{G}$ be a gauge transformation and let $\gamma : [0,1] \to M$ be a curve on M. Let Π_γ and Π_γ^Φ denote the operations of parallel transport along γ with respect to the connections H and H^Φ, respectively. Show that

$$\Phi_{\gamma(1)} \circ \Pi_\gamma = \Pi_\gamma^\Phi \circ \Phi_{\gamma(0)} \,. \tag{23}$$

Suppose now that H is a connection which is fixed by a gauge transformation $\Phi \in \mathscr{G}$. Then for all curves γ, $\Pi_\gamma = \Pi_\gamma^\Phi$, and in particular for all *loops*,

$$\Phi_{\gamma(0)} \circ \Pi_\gamma = \Pi_\gamma \circ \Phi_{\gamma(0)} \,.$$

If the connection is irreducible, then every group element in G is realisable as Π_γ for some loop γ, and the above equation says that $\Phi_{\gamma(0)}$ commutes with all group elements. In other words, it is central and hence trivial in the adjoint group. For example, if $G = SU(2)$ this means that $\Phi_{\gamma(0)}$ is ± 1.

Let $o \in M$ be any point and consider those gauge transformations which are the identity at o. These gauge transformations form a normal subgroup $\mathscr{G}_o \subset \mathscr{G}$, whose quotient $\mathscr{G}/\mathscr{G}_o$ is isomorphic to G. It is not hard to see, again using the gauge covariance of parallel transport, that $\mathscr{G}_o$ acts freely on the space $\mathscr{A}$ of connections. Indeed, suppose that $\Phi \in \mathscr{G}_o$ leaves invariant a connection H. Then again $\Pi_\gamma = \Pi_\gamma^\Phi$ for all curves γ starting at $\gamma(0) = o$, whence using that $\Phi_{\gamma(0)} = \mathrm{id}$,

$$\Phi_{\gamma(1)} \circ \Pi_\gamma = \Pi_\gamma \implies \Phi_{\gamma(1)} = \mathrm{id} \,.$$

Since γ is an arbitrary curve, $\Phi = \mathrm{id}$ everywhere.

In summary, the group of gauge transformations $\mathscr{G}$ acts (almost) freely on the space of irreducible connections and the group of restricted gauge transformations $\mathscr{G}_o$ acts freely on the space of connections. This prompts the following definitions. We will work with definiteness with self-dual connections, but similar definitions apply for anti-self-dual connections.

Let $\mathscr{A}^+ \subset \mathscr{A}$ denote the space of self-dual connections and let $\mathscr{A}_o \subset \mathscr{A}$ denote the space of irreducible connections. Their intersection $\mathscr{A}_o^+ = \mathscr{A}^+ \cap \mathscr{A}_o$ is then the space of irreducible self-dual connections. Both irreducibility and self-duality are gauge invariant conditions, whence $\mathscr{G}$ preserves $\mathscr{A}_o^+$. The quotient $\mathscr{M} = \mathscr{A}_o^+/\mathscr{G}$ is called the **moduli space of instantons**. Alternatively we can consider the quotient $\widetilde{\mathscr{M}} = \mathscr{A}^+/\mathscr{G}_o$, which is called the **moduli space of *framed* instantons**. $\widetilde{\mathscr{M}}$ is fibred over $\mathscr{M}$ with fibres G. Under suitable conditions, both $\mathscr{M}$ and $\widetilde{\mathscr{M}}$ are finite-dimensional manifolds; although it is $\widetilde{\mathscr{M}}$ which has the more interesting geometry, as we will see.

5.2 The deformation complex

Let ω be a self-dual connection. The tangent space $T_\omega\mathscr{A}^+$ is the subspace of $T_\omega\mathscr{A}$ defined by the linearised self-duality equations. In turn, $T_\omega\mathscr{A}^+$ has a subspace consisting of tangent directions to the orbit $\mathscr{G}\cdot\omega$ of ω under gauge transformations. If ω is also irreducible, then the orthogonal complement of $T_\omega(\mathscr{G}\cdot\omega)$ (with respect to a suitable inner product) inside $T_\omega\mathscr{A}^+$ is isomorphic to the tangent space $T_\omega\mathscr{M}$ to the moduli space of instantons at ω. In this section we will set up the calculation of the dimension of $T_\omega\mathscr{M}$. The details can be found in the paper [AHS78].

Since $\mathscr{A}$ is an affine space modelled on $\Omega^1(M;\operatorname{ad}P)$, the tangent space $T_\omega\mathscr{A}$ is naturally isomorphic to $\Omega^1(M;\operatorname{ad}P)$. Consider a curve $\omega_t := \omega + t\tau$ in $\mathscr{A}$ passing through ω, where $\tau \in \Omega^1(M;\operatorname{ad}P)$. Such a straight line will not generally correspond to a self-dual connection for any $t > 0$, but we can demand that it be so up to first order in t; that is, we can demand that its velocity be tangent to $\mathscr{A}^+$. The curvature Ω_t of ω_t is given by

$$\Omega_t = \Omega + t d^\omega \tau + \tfrac{1}{2} t^2 [\tau, \tau] ,$$

where Ω is the curvature of ω. This is self-dual up to first order if and only if $d^\omega\tau$ is self-dual.

In order to recognise those directions tangent to the gauge orbit, we need to discuss infinitesimal gauge transformations. We will consider a curve Φ_t in $\mathscr{G}$ passing by $\Phi_0 = \mathrm{id}$. The derivative with respect to t at the identity gives rise to an element of the tangent space to $\mathscr{G}$ at the identity, which we may identify with Lie algebra $\mathfrak{G} = C^\infty(M; \mathrm{ad}\,P)$ of the group of gauge transformations. We can define a map $\exp : \mathfrak{G} \to \mathscr{G}$ by fibrewise application of the exponential map.[1] We may describe this locally relative to a trivialisation. If $\Theta \in C^\infty(M; \mathrm{ad}\,P)$ is described by a family of local functions $\{\theta_\alpha : U_\alpha \to \mathfrak{g}\}$, then $\Phi_t := \exp(t\Theta) \in \mathscr{G}$ is described by the family of local functions $\{\exp(t\theta_\alpha) : U_\alpha \to G\}$. The connection ω is similarly described by a family of local gauge fields $\{A_\alpha \in \Omega^1(U_\alpha; \mathfrak{g})\}$, on which the gauge transformation Φ_t has the following effect

$$A_\alpha^{\Phi_t} = \exp(t\theta_\alpha) A_\alpha \exp(-t\theta_\alpha) - d\exp(t\theta_\alpha)\exp(-t\theta_\alpha) ,$$

where we have assumed a matrix group for simplicity. Differentiating with respect to t and setting $t = 0$ we recover the form of an infinitesimal gauge transformation:

$$\left.\frac{d}{dt} A_\alpha^{\Phi_t}\right|_{t=0} = \theta_\alpha A_\alpha - A_\alpha \theta_\alpha - d\theta_\alpha = -d_A \theta_\alpha ,$$

which are (up to a sign) the local representatives of $d_A\Theta \in \Omega^1(M; \mathrm{ad}\,P)$.

The preceding discussion can be summarised in terms of the following sequence of linear maps:

$$0 \longrightarrow \Omega^0(M; \mathrm{ad}\,P) \xrightarrow{d_A} \Omega^1(M; \mathrm{ad}\,P) \xrightarrow{d_A^-} \Omega^2_-(M; \mathrm{ad}\,P) \longrightarrow 0 , \tag{24}$$

where $d_A^-\tau := (d_A\tau)^-$ stands for the anti-self-dual part of $d_A\tau$. Notice that because ω is a self-dual connection, the composition $d_A^- \circ d_A = F_A^- = 0$, so the above is a complex, called the **deformation complex**. This means that the image of the first map is contained in the kernel of the second, but it need not necessarily be all of it.

In fact, a tangent vector $\tau \in T_\omega \mathscr{A}$ is tangent to $\mathscr{A}^+$ if and only if it is in the kernel of the second map, whereas it is tangent to the gauge orbit if and only if it is in the image of the first. In other words, if ω is irreducible,

$$T_\omega \mathscr{M} \cong \frac{\ker d_A^- : \Omega^1(M; \mathrm{ad}\,P) \to \Omega^2_-(M; \mathrm{ad}\,P)}{\mathrm{im}\, d_A : \Omega^0(M; \mathrm{ad}\,P) \to \Omega^1(M; \mathrm{ad}\,P)} ,$$

which is the first cohomology group H^1 of the deformation complex. For M compact, the deformation complex is elliptic and hence has finite index

$$\mathrm{index} = \dim H^0 - \dim H^1 + \dim H^2 .$$

In other words,

$$\dim \mathscr{M} = \dim H^1 = -\mathrm{index} + \dim H^0 + \dim H^2 .$$

The index can be computed in principle by the Atiyah–Singer index theorem, but the index will not be enough to compute the dimension of the moduli space unless we have some control over H^0 and H^2.

For an irreducible connection, $\dim H^0 = 0$. Indeed, $H^0 = \ker d_A : \Omega^0(M; \mathrm{ad}\,P) \to \Omega^1(M; \mathrm{ad}\,P)$, hence $\dim H^0 \neq 0$ if and only if there is some $\Theta \in \Omega^0(M; \mathrm{ad}\,P)$ such that $d_A\Theta = 0$. But such a Θ is invariant under parallel transport and hence commutes with the holonomy group of the connection. In particular, it belongs to the centraliser of its Lie algebra. If the connection is irreducible, this Lie algebra is all of $\mathfrak{g}$, which is assumed to be semisimple and hence without centre.

In fact, for the 4-sphere and in the case of $G = SU(2)$, there can be no self-dual reducible connections with nonzero instanton number. The reason is that if the holonomy is a proper subgroup of $SU(2)$, it

[1] Although contrary to what happens in finite-dimensional Lie groups, there may be gauge transformations which are arbitrarily close to the identity which are not in the image of the exponential map.

must have the homotopy type of a circle subgroup and for the 4-sphere, there can be no nontrivial circle bundles, since G-bundles over the 4-sphere are classified by the third homotopy group $\pi_3(G)$, which vanishes for $G = S^1$.

For the 4-sphere (more generally any self-dual manifold with positive scalar curvature), a Weitzenböck argument shows that $H^2 = 0$. This argument runs as follows. H^2 is the cokernel of d_A^-, which is the kernel of the (formal) adjoint $(d_A^-)^*$. One calculates the corresponding laplacian operator $d_A(d_A^-)^*$ and shows that this is a positive operator and hence that it has no kernel.

Therefore on the 4-sphere, $\dim \mathscr{M}$ coincides with the index of the deformation complex, which can be computed using the Atiyah–Singer index theorem. For gauge group $G = SU(2)$ and instanton number k (positive), one obtains $\dim M = 8k - 3$. In particular, for $k = 1$ we obtain a five-dimensional moduli space. These are precisely the five parameters in the BPST solution: the scale and the centre of the instanton.

(I realise that this section is missing many details. I hope to remedy this eventually by a couple of lectures on a supersymmetric proof of the index theorem.)