3

Interlude on sets

Sets and functions are ubiquitous in mathematics. You might have the impression that they are most strongly connected with the pure end of the subject, but this is an illusion: think of probability density functions in statistics, data sets in experimental science, planetary motion in astronomy, or flow in fluid dynamics.

Category theory is often used to shed light on common constructions and patterns in mathematics. If we hope to do this in an advanced context, we must begin by settling the basic notions of set and function. That is the purpose of the first section of this chapter.

The definition of category mentions a 'collection' of objects and 'collections' of maps. We will see in the second section that some collections are too big to be sets, which leads to a distinction between 'small' and 'large' collections. This distinction will be needed later, most prominently for the adjoint functor theorems (Chapter 6).

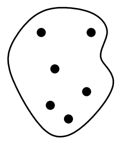
The final section takes a historical look at set theory. It also explains why the approach to sets taken in this chapter is more relevant to most of mathematics than the traditional approach is. None of this section is logically necessary for anything that follows, but it may provide useful perspective.

I do not assume that you have encountered axiomatic set theory of any kind. If you have, it is probably best to put it out of your mind while reading this chapter, as the approach to set theory that we take is quite different from the approach that you are most likely to be familiar with. A brief comparison of the traditional and categorical approaches can be found at the very end of the chapter.

3.1 Constructions with sets

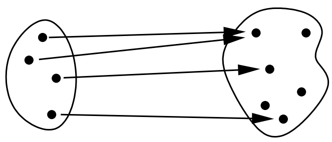
We have made no definition of ‘set’, nor of ‘function’. Nevertheless, guided by our intuition, we can list some properties that we expect the world of sets and functions to have. For instance, we can describe some of the sets that we think ought to exist, and some ways of building new sets from old.

Intuitively, a set is a bag of points:

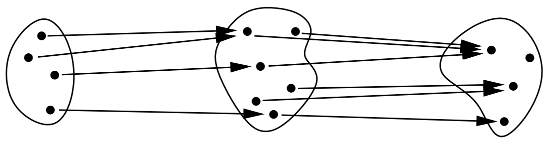


(There may, of course, be infinitely many.) These points, or elements, are not related to one another in any way. They are not in any order, they do not come with any algebraic structure (for instance, there is no specified way of multiplying elements together), and there is no sense of what it means for one point to be close to another. In particular examples, we might have some extra structure in mind; for instance, we often equip the set of real numbers with an order, a field structure and a metric. But to view $\mathbb{R}$ as a mere *set* is to ignore all that structure, to regard it as no more than a bunch of featureless points.

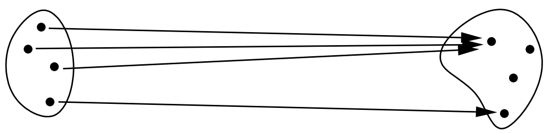
Intuitively, a function $A \to B$ is an assignment of a point in bag B to each point in bag A:



We can do one function after another: given functions



we obtain a composite function



This composition of functions is associative: $h \circ (g \circ f) = (h \circ g) \circ f$. There is also an identity function on every set. Hence:

Sets and functions form a category, denoted by **Set***.*

This does not pin things down much: there are many categories, mostly quite unlike the category of sets. So, let us list some of the special features of the category of sets.

The empty set There is a set $\emptyset$ with no elements.

Suppose that someone hands you a pair of sets, A and B, and tells you to specify a function from A to B. Then your task is to specify for each element of A an element of B. The larger A is, the longer the task; the smaller A is, the shorter the task. In particular, if A is empty then the task takes no time at all; we have nothing to do. So there is a function from $\emptyset$ to B specified by doing nothing. On the other hand, there cannot be two different ways to do nothing, so there is only one function from $\emptyset$ to B. Hence:

$\emptyset$ *is an initial object of* **Set***.*

In case this argument seems unconvincing, here is an alternative. Suppose that we have a set A with disjoint subsets A_1 and A_2 such that $A_1 \cup A_2 = A$. Then a function from A to B amounts to a function from A_1 to B together with a function from A_2 to B. So if all the sets are finite, we should have the rule

$$\begin{aligned}(\text{number of functions from } A \text{ to } B) = {} & (\text{number of functions from } A_1 \text{ to } B) \\ & \times (\text{number of functions from } A_2 \text{ to } B).\end{aligned}$$

In particular, we could take $A_1 = A$ and $A_2 = \emptyset$. This would force the number of functions from $\emptyset$ to B to be 1. So if we want this rule to hold (and surely we do!), we had better say that there is exactly one function from $\emptyset$ to B.

What about functions *into* $\emptyset$? There is exactly one function $\emptyset \to \emptyset$, namely, the identity. This is a special case of the initiality of $\emptyset$. On the other hand, for a set A that is not empty, there are no functions $A \to \emptyset$, because there is nowhere for elements of A to go.

The one-element set There is a set 1 with exactly one element.

For any set A, there is exactly one function from A to 1, since every element of A must be mapped to the unique element of 1. That is:

1 *is a terminal object of* **Set***.*

A function *from* 1 *to* a set B is just a choice of an element of B. In short, the functions $1 \to B$ are the elements of B. Hence:

The concept of element is a special case of the concept of function.

Products Any two sets A and B have a product, $A \times B$. Its elements are the ordered pairs (a, b) with $a \in A$ and $b \in B$. Ordered pairs are familiar from coordinate geometry. All that matters about them is that for $a, a' \in A$ and $b, b' \in B$,

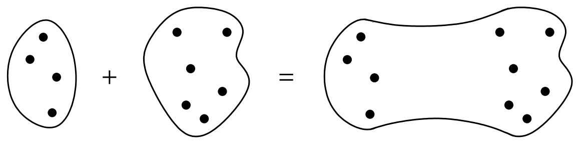
$$(a, b) = (a', b') \iff a = a' \text{ and } b = b'.$$

More generally, take any set I and any family $(A_i)_{i \in I}$ of sets. There is a product set $\prod_{i \in I} A_i$, whose elements are families $(a_i)_{i \in I}$ with $a_i \in A_i$ for each i. Just as for ordered pairs,

$$(a_i)_{i \in I} = (a'_i)_{i \in I} \iff a_i = a'_i \text{ for all } i \in I.$$

Sums Any two sets A and B have a **sum** $A + B$.

Thinking of sets as bags of points, the sum of two sets is obtained by putting all the points into one big bag:



If A and B are finite sets with m and n elements respectively, then $A + B$ always has $m + n$ elements. It makes no difference what the elements of $A + B$ are called; as usual, we only care what $A + B$ is up to isomorphism.

There are inclusion functions

$$A \xrightarrow{i} A + B \xleftarrow{j} B$$

such that the union of the images of i and j is all of $A + B$ and the intersection of the images is empty.

Sum is sometimes called **disjoint union** and written as $\amalg$. It is not to be confused with (ordinary) union $\cup$. For a start, we can take the sum of *any* two sets A and B, whereas $A \cup B$ only really makes sense when A and B come as subsets of some larger set. (For to say what $A \cup B$ is, we need to know which elements of A are equal to which elements of B.) And even if A and B do come as subsets of some larger set, $A + B$ and $A \cup B$ can be different. For example, take the subsets $A = \{1, 2, 3\}$ and $B = \{3, 4\}$ of $\mathbb{N}$. Then $A \cup B$ has 4 elements, but $A + B$ has $3 + 2 = 5$ elements.

More generally, any family $(A_i)_{i \in I}$ of sets has a sum $\sum_{i \in I} A_i$. If I is finite and each A_i is finite, say with m_i elements, then $\sum_{i \in I} A_i$ has $\sum_{i \in I} m_i$ elements.

Sets of functions For any two sets A and B, we can form the set A^B of functions from B to A.

This is a special case of the product construction: A^B is the product $\prod_{b \in B} A$ of the constant family $(A)_{b \in B}$. Indeed, an element of $\prod_{b \in B} A$ is a family $(a_b)_{b \in B}$ consisting of one element $a_b \in A$ for each $b \in B$; in other words, it is a function $B \to A$.

Digression on arithmetic We are using notation reminiscent of arithmetic: $A \times B$, $A + B$, and A^B. There is good reason for this: if A is a finite set with m elements and B a finite set with n elements, then $A \times B$ has $m \times n$ elements, $A + B$ has $m + n$ elements, and A^B has m^n elements. Our notation 1 for a one-element set and the alternative notation 0 for the empty set $\emptyset$ also follow this pattern.

All the usual laws of arithmetic have their counterparts for sets:

$$\begin{aligned} A \times (B + C) &\cong (A \times B) + (A \times C), \\ A^{B+C} &\cong A^B \times A^C, \\ (A^B)^C &\cong A^{B \times C}, \end{aligned}$$

and so on, where $\cong$ is isomorphism in the category of sets. (For the last one, see Example 2.1.6.) These isomorphisms hold for all sets, not just finite ones.

The two-element set Let 2 be the set 1 + 1 (a set with two elements!). For reasons that will soon become clear, I will write the elements of 2 as `true` and `false`.

Let A be a set. Given a subset S of A, we obtain a function $\chi_S : A \to 2$ (the **characteristic function** of $S \subseteq A$), where

$$\chi_S(a) = \begin{cases} \texttt{true} & \text{if } a \in S, \\ \texttt{false} & \text{if } a \notin S \end{cases}$$

($a \in A$). Conversely, given a function $f : A \to 2$, we obtain a subset

$$f^{-1}\{\texttt{true}\} = \{a \in A \mid f(a) = \texttt{true}\}$$

of A. These two processes are mutually inverse; that is, χ_S is the unique function $f : A \to 2$ such that $f^{-1}\{\texttt{true}\} = S$. Hence:

Subsets of A correspond one-to-one with functions $A \to 2$.

We already know that the functions from A to 2 form a set, 2^A. When we are thinking of 2^A as the set of all subsets of A, we call it the **power set** of A and write it as $\mathscr{P}(A)$.

Equalizers It would be nice if, given a set A, we could define a subset S of A by specifying a property that the elements of S are to satisfy:

$$S = \{a \in A \mid \text{some property of } a \text{ holds}\}.$$

It is hard to give a general definition of 'property'. There is, however, a special type of property that is easy to handle: equality of two functions. Precisely, given sets and functions $A \overset{f}{\underset{g}{\rightrightarrows}} B$, there is a set

$$\{a \in A \mid f(a) = g(a)\}.$$

This set is called the **equalizer** of f and g, since it is the part of A on which the two functions are equal.

Quotients You are probably familiar with quotient groups and quotient rings (sometimes called factor groups and factor rings) in algebra. Quotients also come up everywhere in topology, such as when we glue together opposite sides of a square to make a cylinder. But the most basic context for quotients is that of sets.

Let A be a set and $\sim$ an equivalence relation on A. There is a set $A/\sim$, the **quotient of** A **by** $\sim$, whose elements are the equivalence classes. For example, given a group G and a normal subgroup N, define an equivalence relation $\sim$ on G by $g \sim h \iff gh^{-1} \in N$; then $G/\sim = G/N$.

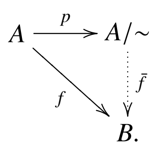
There is also a canonical map

$$p \colon A \to A/\sim,$$

sending an element of A to its equivalence class. It is surjective, and has the property that $p(a) = p(a') \iff a \sim a'$. In fact, it has a universal property: any function $f \colon A \to B$ such that

$$\forall a, a' \in A, \qquad a \sim a' \implies f(a) = f(a') \tag{3.1}$$

factorizes uniquely through p, as in the diagram



Thus, for any set B, the functions $A/\sim \to B$ correspond one-to-one with the functions $f \colon A \to B$ satisfying (3.1). This fact is at the heart of the famous isomorphism theorems of algebra.

We have now listed the properties of sets and functions that will be most important for us. Here are two more.

Natural numbers A function with domain $\mathbb{N}$ is usually called a **sequence**. A crucial property of $\mathbb{N}$ is that sequences can be defined recursively: given a set X, an element $a \in X$, and a function $r\colon X \to X$, there is a unique sequence $(x_n)_{n=0}^{\infty}$ of elements of X such that

$$x_0 = a, \qquad x_{n+1} = r(x_n) \text{ for all } n \in \mathbb{N}.$$

This property refers to two pieces of structure on $\mathbb{N}$: the element 0, and the function $s\colon \mathbb{N} \to \mathbb{N}$ defined by $s(n) = n+1$. Reformulated in terms of functions, and writing $x_n = x(n)$, the property is this: for any set X, element $a \in X$, and function $r\colon X \to X$, there is a unique function $x\colon \mathbb{N} \to X$ such that $x(0) = 0$ and $x \circ s = r \circ x$. Exercise 3.1.2 asks you to show that this is a *universal* property of $\mathbb{N}$, 0 and s.

Choice Let $f\colon A \to B$ be a map in a category $\mathscr{A}$. A **section** (or **right inverse**) of f is a map $i\colon B \to A$ in $\mathscr{A}$ such that $f \circ i = 1_B$.

In the category of sets, any map with a section is certainly surjective. The converse statement is called the **axiom of choice**:

Every surjection has a section.

It is called 'choice' because specifying a section of $f\colon A \to B$ amounts to choosing, for each $b \in B$, an element of the nonempty set $\{a \in A \mid f(a) = b\}$.

The properties listed above are not theorems, since we do not have rigorous definitions of set and function. What, then, is their status?

Definitions in mathematics usually depend on previous definitions. A vector space is defined as an abelian group with a scalar multiplication. An abelian group is defined as a group with a certain property. A group is defined as a set with certain extra structure. A set is defined as... well, what?

We cannot keep going back indefinitely, otherwise we quite literally would not know what we were talking about. We have to start somewhere. In other words, there have to be some basic concepts not defined in terms of anything else. The concept of set is usually taken to be one of the basic ones, which is why you have probably never read a sentence beginning 'Definition: A **set** is...'. We will treat function as a basic concept, too.

But now there seems to be a problem. If these basic concepts are not defined in terms of anything else, how are we to know what they really are? How are we going to reason in the watertight, logical way upon which mathematics depends? We cannot simply trust our intuitions, since your intuitive idea of set might be slightly different from mine, and if it came to a dispute about how sets behave, we would have no way of deciding who was right.

The problem is solved as follows. Instead of *defining* a set to be a such-and-such and a function to be a such-and-such else, we list some *properties* that we assume sets and functions to have. In other words, we never attempt to say what sets and functions *are*; we just say what you can *do* with them.

In his excellent book *Mathematics: A Very Short Introduction*, Timothy Gowers (2002) considers the question: 'What is the black king in chess?' He swiftly points out that this question is rather peculiar. It is not important that the black king *is* a small piece of wood, painted a certain colour and carved into a certain shape. We could equally well use a scrap of paper with 'BK' written on it. What matters is what the black king *does*: it can move in certain ways but not others, according to the rules of chess.

Similarly, we might not be able to say directly what a set or function 'is', but we agree that they are to satisfy all the properties on the list. So the list of properties acts as an agreement on how to use the words 'set' and 'function', just as the rules of chess act as an agreement on how to use the chess pieces.

What we are doing is often referred to as *foundations*. In this metaphor, the foundation consists of the basic concepts (set and function), which are not built on anything else, but are assumed to satisfy a stated list of properties. On top of the foundations are built some basic definitions and theorems. On top of those are built further definitions and theorems, and so on, towering upwards.

The properties above are stated informally, but they can be formalized using some categorical language. (See Lawvere and Rosebrugh (2003) or Leinster (2014).) In the formal version, we begin by saying that sets and functions form a category, **Set**. We then list some properties of this category. For example, the category is required to have an initial and a terminal object, and the properties described informally under the headings 'Products' and 'Equalizers' are made formal by the statement that **Set** 'has limits' (a phrase defined in Chapter 5).

While we were making the list, we were guided by our intuition about sets. But once it is made, our intuition plays no further official role: any disputes about the nature of sets are settled by consulting the list of properties.

(A subtlety arises. Whatever list of properties one writes down, there might be some questions that cannot be settled. In other words, there might be multiple inequivalent categories satisfying all the properties listed. This gets us into the realm of advanced logic: Gödel incompleteness, the continuum hypothesis, and so on, all beyond the scope of this book.)

Now let us look again at the section on the empty set. You might have felt that I was on shaky ground when trying to convince you that $\emptyset$ is initial. But the point is that I do not need to convince you that this is a *true statement*; I only need to convince you that it is a *convenient assumption*. Compare the rule for numbers that $x^0 = 1$. One can reasonably argue that 0 copies of x multiplied together ought to be 1, but really the best justification for this rule is convenience: it makes other rules such as $x^{m+n} = x^m \cdot x^n$ true without exception. Indeed, it does not even make sense to ask whether it is 'true' that $\emptyset$ is initial until we have written down our assumptions about how sets and functions behave. For until then, what could 'true' mean? There is no physical world of sets against which to test such statements.

We can make whatever assumptions about sets we like, but some lead to more interesting mathematics than others. If, for instance, you want to assume that there are *no* functions from $\emptyset$ to any other set, you can, but the tower of mathematics built on that foundation will look different from what you are used to, and probably not in a good way. For example, the 'number of functions' rule (page 67) will fail, and there will be further unpleasant surprises higher up the tower.

Exercises

3.1.1 The **diagonal functor** $\Delta\colon \mathbf{Set} \to \mathbf{Set} \times \mathbf{Set}$ is defined by $\Delta(A) = (A, A)$ for all sets A. Exhibit left and right adjoints to Δ.

3.1.2 In the paragraph headed 'Natural numbers', it was observed that the set $\mathbb{N}$, together with the element 0 and the function $s\colon \mathbb{N} \to \mathbb{N}$, has a certain property. This property can be understood as stating that the triple $(\mathbb{N}, 0, s)$ is the initial object of a certain category $\mathscr{C}$. Find $\mathscr{C}$.

3.2 Small and large categories

We have now made some assumptions about the nature of sets. One consequence of those assumptions is that in many of the categories we have met, the collection of all objects is too large to form a set. In fact, even the collection of *isomorphism classes* of objects is often too large to form a set. In this section, I will explain what these statements mean, and prove them.

This section is not of central importance. As this book proceeds, I will say as little as possible about the distinction between sets and collections too large to be sets. Nevertheless, the distinction begins to matter in parts of category theory lying just within the scope of this book (the adjoint functor theorems), as well as beyond.