Foundations

2

2.1 Objects, Relations, and Operations

2.1.1 Distinctions

A *set* is a collection of distinct or distinguishable objects, its elements. But how can these elements be distinguished? Possibly, by certain specific intrinsic properties that they possess in contrast to others. Better, by specific relations that they have with other elements.

This leads to the concept of *equivalence*, of "equal" versus "different", or of "nondistinguishable" versus "distinguishable". Objects that have the same properties or that stand in the same relation to all other elements are equivalent, as they cannot be distinguished from each other. Therefore, one might want to identify them, that is, treat them as the same and not as different objects. (One should be aware of the possibility, however, that the identification need not be unique because the objects may possess certain internal symmetries or automorphisms, a concept to be explained below.)

So, when we have a set of points, we may not be able to distinguish these points from each other, and therefore, we should identify them all. The resulting set would then consist of a single point only. However, strangely, we can distinguish different sets by their number of points or elements. That is, as soon as we can distinguish the elements of a set, we can also distinguish different sets. As we shall see, however, two sets with the same number of elements cannot be distinguished from each other, unless we can distinguish the elements themselves between the two sets.

Returning to the intrinsic properties, an element could have some internal structure, for instance be a set itself, or some space (see below), that is, it could be an object with structure. Or, conversely, we could say that an object consists of elements with relations between them. In any case, however, this seems to be some kind of higher level element. But the basic mathematical formalism ignores this kind of hierarchy. A collection of sets can again be treated as a set (so long as certain paradoxes of self-reference are avoided). More abstractly, we shall introduce the notion of a category below.

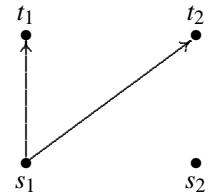
J. Jost, *Mathematical Concepts*, DOI 10.1007/978-3-319-20436-9_2

2.1.2 Mappings

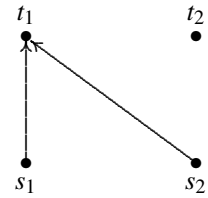
In this section, we start with an elementary concept that most likely will be familiar to our readers. We utilize this also to introduce diagrams as a useful tool to visualize examples for some of the general concepts that we shall introduce in this chapter, and perhaps also to compensate our readers a little for the steepness of some of the subsequent conceptual developments.

Let us consider two sets S and T. We consider a special type of relation between their elements.

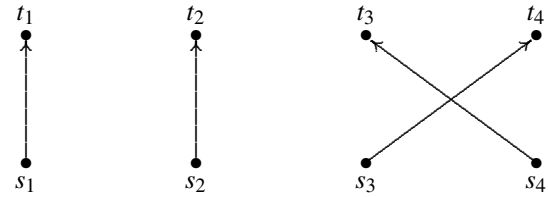
Definition 2.1.1 A *map*, also called a *mapping*, $g : S \to T$ assigns to each element $s \in S$ one and only one element $g(s) \in T$, also written as $s \mapsto g(s)$. Such a map $g : S \to T$ between sets is called *injective* if whenever $s_1 \neq s_2 \in S$, then also $g(s_1) \neq g(s_2)$. That is, different elements in S should have different images in T. The map $g : S \to T$ is called *surjective* if for every $t \in T$, there exists some, in general not unique, $s \in S$ with $g(s) = t$. Thus, no point in T is omitted from the image of S. Of course, if g were not surjective, we could simply replace T by $g(S)$ to make it surjective. A map $g : S \to T$ that is both injective and surjective is called *bijective*.



This does not define a map from the set $\{s_1, s_2\}$ to the set $\{t_1, t_2\}$ because s_1 has two images, instead of only one, whereas s_2 has none.



This is a map, but it is neither injective, because t_1 has two different preimages, s_1 as well as s_2, nor surjective, because t_2 has no preimage at all.



This represents a bijective map from $\{s_1, s_2, s_3, s_4\}$ to $\{t_1, t_2, t_3, t_4\}$.

Let us consider the set $\{1\}$ consisting of a single element only. Let S be another non-empty set. We can then specify an element s of S by a map

$$f : \{1\} \to S \text{ with } f(1) = s. \tag{2.1.1}$$

Thus, the set S corresponds to the maps $f : \{1\} \to S$. The set $\{1\}$ serves as a universal spotlight that can be directed by a map f to any particular element of the set S.

When S' is a subset of S, $S' \subset S$, then we have the inclusion map

$$\begin{aligned} i : S' &\to S \\ s &\mapsto s \text{ for every } s \in S'. \end{aligned} \tag{2.1.2}$$

This map i is injective, but not surjective, unless $S' = S$.

More generally, a binary relation between the sets S and T is given by a collection of ordered pairs $R = \{(s, t)\}$ where $s \in S, t \in T$. While a relation is more general than a map, we can represent every such relation R by a map

$$\begin{aligned} r : S \times T &\to \{0, 1\} \\ r((s, t)) = 1 \quad &\Leftrightarrow \quad (s, t) \in R, \end{aligned} \tag{2.1.3}$$

and hence, equivalently, $r((s, t)) = 0$ iff $(s, t) \notin R$. We might say here that 1 stands for "true", that is, the relation R holds, whereas 0 stands for "false", that is, R does not hold for the pair (s, t).

Mappings can be composed. That is, if $f : S \to T$ and $g : T \to V$ are maps, then the map $h := g \circ f$ is defined by

$$s \mapsto f(s) \mapsto g(f(s)), \tag{2.1.4}$$

i.e., s is mapped to the image of $f(s)$ under the map g, $g(f(s))$. We note that for this procedure to be possible, the target T of f, also called its *codomain*, has to agree with the *domain* of the map g.

Lemma 2.1.1 *The composition of maps is associative. This means that when $f : S \to T, g : T \to V, h : V \to W$ are maps, then*

$$h \circ (g \circ f) = (h \circ g) \circ f =: h \circ g \circ f. \tag{2.1.5}$$

Proof Under either of the variants given in (2.1.5), the image of $s \in S$ is the same, $h(g(f(s)))$. □

Let us also explain the use of the brackets (.). The expression $h \circ (g \circ f)$ means that we first compute the composition $g \circ f$—let us call it η, and then the composition $h \circ \eta$. For $(h \circ g) \circ f$, it is the other way around. We first compute the composition $h \circ g =: \phi$ and then $\phi \circ f$. (2.1.5) tells us that the two results agree. In general, brackets (.) are used to specify the order in which the various operations in a formula have to be carried out. In some cases, there exists a general convention for such an order, and in those cases, brackets will not be needed. For instance, in an expression $a \cdot b + c$, we first compute the product $a \cdot b$ and then add c to it. Also, when two expressions are connected by an equality or inequality sign, like $a + b \leq c \cdot d$, then the expressions on the left- and on the right-hand side of that sign are each first

computed and then the corresponding results are compared. But probably, every reader knows that. We shall repeatedly make use of such implicit conventions.

2.1.3 Power Sets and Distinctions

We consider a set S with finitely many elements, say $S = \{s_1, s_2, \ldots, s_n\}$. And we assume that there is some property that may either apply or not to each element of S. We write $P(s) = 1$ when s satisfies this property, and $P(s) = 0$ if it does not. According to the separation principle of Zermelo-Frankel set theory (see Sect. 2.2), each such P specifies the subset PS of those s that satisfy $P(s) = 1$, i.e.,

$$PS = \{s \in S : P(s) = 1\}. \tag{2.1.6}$$

Conversely, for any subset S' of S, we can define such a property $P_{S'}$ by

$$P_{S'}(s) = 1 \text{ if and only if } s \in S'. \tag{2.1.7}$$

Thus,

$$P_{S'}S = S'. \tag{2.1.8}$$

We call the set of all subsets of S its *power set* $\mathcal{P}(S)$. Thus, each *subset* $S' \subset S$ becomes an *element* $S' \in \mathcal{P}(S)$. We can also say that each element of $\mathcal{P}(S)$ corresponds to a *distinction* that we can make on the elements of S, whether they possess a property P or not.

We can also interpret P as a map

$$P : S \to \{0, 1\} \tag{2.1.9}$$

from S into the set $\boldsymbol{2} := \{0, 1\}$, and we shall therefore also write

$$\mathcal{P}(S) = \boldsymbol{2}^S. \tag{2.1.10}$$

Now, when S is the empty set $\emptyset$, then its power set, the set of all its subsets, is

$$\mathcal{P}(\emptyset) = \{\emptyset\}, \tag{2.1.11}$$

because the empty set is a subset of every set, hence also of itself. Thus, the power set of the empty set is not empty, but contains $\emptyset$ as its single element. That is, the power set of the empty set contains one element. This is the trivial distinction, but it is a distinction nonetheless.

$\{1\}$

Next, when $S = \{1\}$ contains a single element, which we now denote by 1, then its power set is

$$\mathcal{P}(\{1\}) = \{\emptyset, \{1\}\}, \tag{2.1.12}$$

because we now have two possible properties or distinctions. When 1 does not satisfy the property, we get $PS = \emptyset$, but when 1 does satisfy it, we have $PS = \{1\}$.

$\{1, 2\}$

Moving on to a set $S = \{1, 2\}$ with two elements, we have

$$\mathcal{P}(\{1, 2\}) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}, \tag{2.1.13}$$

because now a property can be satisfied by none, both or either of the elements. That is, on two elements, we can make four different distinctions.

Pursuing this pattern, we see that the power set of a set with n elements has 2^n elements. Thus, the size of the power set grows exponentially as a function of the size of the original set.

We next turn to infinite sets. We start with the positive integers $\mathbb{N} = \{1, 2, 3, \dots\}$ and call a set S *infinite* if there exists an *injective map*

$$i : \mathbb{N} \to S. \tag{2.1.14}$$

$\mathbb{N}$ possesses infinite subsets that do not contain all the elements of $\mathbb{N}$ itself. For instance, the set of even positive integers, $2\mathbb{N} = \{2, 4, 6, \dots\}$, is infinite because we have the injective map $i : \mathbb{N} \to 2\mathbb{N}$ with $i(n) = 2n$ for $n \in N$. In fact, in this case, the map i is even bijective. There exist other sets S which may appear larger than $\mathbb{N}$ from a naive perspective for which we nevertheless have a bijection $i : \mathbb{N} \to S$. For example, let us consider the set $S = \{(n, m)\}$ of all pairs of positive integers n, m. We construct a bijection by the following procedure:

$$\begin{aligned} &1 \mapsto (1, 1), \\ &2 \mapsto (1, 2), \quad 3 \mapsto (2, 1), \\ &4 \mapsto (1, 3), \quad 5 \mapsto (2, 2), \quad 6 \mapsto (3, 1), \\ &7 \mapsto (1, 4), \dots, \end{aligned}$$

that is, for every $k \in \mathbb{N}$ we enumerate the finitely many pairs with $n+m = k$ and after that move on to $k + 1$. Similarly, we can construct a bijection between $\mathbb{N}$ and the set of all N-tuples of elements of $\mathbb{N}$, for every $N \in \mathbb{N}$.

However, this is not possible between $\mathbb{N}$ and its power set $\mathcal{P}(\mathbb{N})$. Every element of $X \in \mathcal{P}(\mathbb{N})$ corresponds to a distinction that we can make in $\mathbb{N}$, that is, to a property that we can check for every $n \in \mathbb{N}$ and then assign those n that satisfy it to the set X. We can also express this via a binary sequence, like

$$100110100\dots \tag{2.1.15}$$

which means that the integers $1, 4, 5, 7, \dots$ satisfy the property whereas $2, 3, 6, 8, 9, \dots$ don't. We now describe Cantor's famous diagonal argument that the set of all such binary sequences $\sigma = (\sigma_1, \sigma_2, \sigma_3, \dots)$ cannot be put in a bijective correspondence with $\mathbb{N}$. The argument proceeds by contradiction. Suppose that there were such a correspondence, $i : n \mapsto \sigma(n)$. We then consider the sequence σ' constructed as follows. When $\sigma_k(k) = 1$, we put $\sigma'(k) = 0$, and if $\sigma_k(k) = 0$, we put $\sigma'(k) = 1$. Thus, at the kth position, σ' is different from $\sigma(k)$. Therefore, for every k, there is some position for which σ' is different from $\sigma(k)$. Thus, σ' is different from all the sequences $\sigma(k)$. But that means that the correspondence i is not surjective, and this is the contradiction.

In general, the power set $\mathcal{P}(S)$ of a set S is always "larger" than S itself, in the sense that on a set of elements, we can make more distinctions than there are elements.

Cantor's argument showed that there is no surjective map

$$i : \mathbb{N} \to \boldsymbol{2}^{\mathbb{N}}. \tag{2.1.16}$$

We shall now generalize the argument and the result (and the reader may want to skip the rest of this section upon a first reading) and show that only under very special circumstances can there be a surjective map

$$\begin{aligned} f &: S \to \Lambda^S \\ x &\mapsto f_x : S \to \Lambda \end{aligned} \tag{2.1.17}$$

where

$$\Lambda^S := \{\lambda : S \to \Lambda\} \tag{2.1.18}$$

is the set of all maps from S to Λ. Each map $f : S \to \Lambda^S$ (whether surjective or not) yields

$$\begin{aligned} \tilde{f} &: S \times S \to \Lambda \\ \tilde{f}(x, y) &= f_x(y), \end{aligned} \tag{2.1.19}$$

that is, for x, we have the map $f_x : S \to \Lambda$ which we can apply to $y \in S$. We then have

Lemma 2.1.2 *If there is a surjective map*

$$g : S \to \Lambda^S \tag{2.1.20}$$

then every map

$$\lambda : \Lambda \to \Lambda \tag{2.1.21}$$

has a fixed point, that is, there exists some $\ell \in \Lambda$ with

$$\lambda(\ell) = \ell. \tag{2.1.22}$$

Proof We consider the diagonal embedding

$$\begin{aligned} \Delta &: S \to S \times S \\ x &\mapsto (x, x) \end{aligned} \tag{2.1.23}$$

and, recalling (2.1.19), the map

$$\phi := \lambda \circ \tilde{g} \circ \Delta : S \to \Lambda. \tag{2.1.24}$$

Anticipating the notations of category theory to be introduced in Sect. 2.3, we can represent this by a diagram

$$\begin{array}{ccc} S \times S & \xrightarrow{\tilde{g}} & \Lambda \\ \Delta \uparrow & & \downarrow \lambda \\ S & \xrightarrow{\phi} & \Lambda \end{array} \quad . \tag{2.1.25}$$

Now, if g is surjective, there has to be some $x_0 \in S$ with

$$g(x_0) = \phi, \text{ or equivalently, } \tilde{g}(x_0, y) = \phi(y) \text{ for all } y \in S, \tag{2.1.26}$$

and hence in particular, and this is the crucial diagonal argument,

$$\tilde{g}(x_0, x_0) = \phi(x_0). \tag{2.1.27}$$

But then

$$\phi(x_0) = \lambda \circ \tilde{g} \circ \Delta(x_0) = \lambda \circ \tilde{g}(x_0, x_0) = \lambda(\phi(x_0)),$$

that is,

$$\ell = \phi(x_0)$$

satisfies (2.1.22), i.e., is a fixed point. □

Now, of course, for $\Lambda = \{0, 1\}$, the map λ with $\lambda(0) = 1, \lambda(1) = 0$ does not have a fixed point. Therefore, there can be no surjective map $g : S \to \{0, 1\}$ for any S, and in particular not for $S = \mathbb{N}$. Thus, Cantor's result holds. Note that Cantor's idea is translated into the diagonal operator Δ and the formula (2.1.27) in the above proof.

$\{0, 1\}$

More generally, on any set Λ with more than one element, we can permute the elements to construct a map without fixed points. Therefore, the above argument can be translated into a proof by contradiction. For any set Λ with more than one (for instance two) elements, the existence of a surjective map (2.1.20) would lead to a contradiction.

2.1.4 Structures

2.1.4.1 Binary Relations

We now look at a single set S. A *structure* consists of relations between the elements of the set S. Often, these relations are conceived or imagined as spatial relations. This leads to the concept of a *space*, to be defined below. This brings us into the realm of *geometry*.

A *relation* on a set S is given by a map

$$F : S \times S \to R \tag{2.1.28}$$

for some range set R. (In Chap. 3, we shall also consider relations involving more than two elements of a set S.)

Definition 2.1.2 When we have two sets with relations, (S_1, F_1) and (S_2, F_2), with the same range R, then we call a map $\phi : S_1 \to S_2$ a *homomorphism* (i.e., structure preserving) if for all $s, s' \in S_1$

$$F_1(s, s') = r \text{ implies } F_2(\phi(s), \phi(s')) = r \tag{2.1.29}$$

for all $r \in R$. We shall then also write this as $\phi : (S_1, F_1) \to (S_2, F_2)$.

Definition 2.1.3 Let $F : S \times S \to R$ be a relation and $\phi : S' \to S$ a map. We then define the *pullback relation* $\phi^* F : S' \times S' \to R$ by

$$\phi^* F(s', s'') = F(\phi(s'), \phi(s'')) \text{ for } s', s'' \in S'. \tag{2.1.30}$$

In particular, when $S' \subset S$, we can pull back a relation from S to S' via the inclusion (2.1.2).

With this definition,

$$\phi : (S', \phi^* F) \to (S, F) \tag{2.1.31}$$

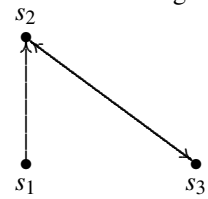
becomes a homomorphism. We record this observation as a principle.

Theorem 2.1.1 *Relations can be pulled back by mappings, and the corresponding mappings then become homomorphisms.*

The simplest relation is a binary one, as explained at the end of Sect. 2.1.2. That is, two elements either stand in a relation, or they don't. When S is our set, according to (2.1.3), this can be expressed as

$$F : S \times S \to \{0, 1\}. \tag{2.1.32}$$

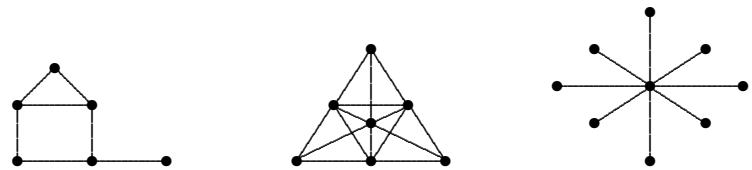
This is also known as a *directed graph* (sometimes also called a *digraph*) with vertex set S and with an ordered pair $(s_1, s_2) \in S \times S$ being an edge when $F(s_1, s_2) = 1$. We also call this an edge from s_1 to s_2.



This depicts a digraph with edges (s_1, s_2), (s_2, s_3), (s_3, s_2).

When F is *symmetric*, that is $F(s_1, s_2) = F(s_2, s_1)$ for all s_1, s_2, then this yields an undirected graph, usually simply called a *graph* for short. Here are some pictures of graphs.

Some graphs



The second and the third are highly symmetric whereas the first does not exhibit any symmetries. Here, a *symmetry* would be a bijective homomorphism h from the vertex set to itself. That it is a homomorphism refers to the edge relation. It means that (s_1, s_2) is an edge precisely if $(h(s_1), h(s_2))$ is. Anticipating our discussion of automorphism groups, we see that such symmetries can be composed, that is, if h_1 and h_2 are symmetries, then so is $h_2 \circ h_1$. As an exercise, the reader might determine all the symmetries of the latter two graphs and their composition rules. The answer for the last graph is that the vertex in the center has to stay fixed under any symmetry whereas any permutation of the other 8 vertices yields a symmetry.

We now introduce some important types of relations.

Definition 2.1.4 When the binary relation F is

reflexive: $F(s, s) = 1$ for all s

transitive: if $F(s_1, s_2) = 1$ and $F(s_2, s_3) = 1$, then also $F(s_1, s_3) = 1$

symmetric: $F(s_1, s_2) = F(s_2, s_1)$,

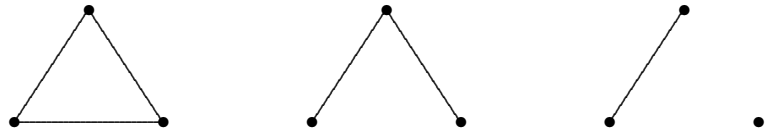
then we say that F defines an *equivalence relation* on S. In this case, one usually writes $s_1 \equiv s_2$ for $F(s_1, s_2) = 1$. We then define a quotient S/F of

S by the equivalence relation F whose elements are the equivalence classes

$$[s] := \{s' \in S : s' \equiv s\} \tag{2.1.33}$$

for $s \in S$.

In this section, we shall draw a number of graphs or digraphs to illustrate algebraic structures geometrically. We shall usually omit edges from the vertices to themselves, to simplify our drawings. That is, when depicting reflexive relations, the reflexivity condition is always assumed, but not drawn.



Here, two vertices s, s' are connected by an edge iff $F(s, s') = 1$. The first and the third graph here then represent equivalence relations, whereas the second one does not, as it is not transitive.

One can view an equivalence relation F on S as a partitioning of S into the equivalence classes. In the quotient S/F, equivalence is transformed into equality, $[s_1] = [s_2]$ iff $s_1 \equiv s_2$. We also obtain an induced relation F_q on S/F by putting $F_q([s], [s']) = 1$ if $[s] = [s']$ and $= 0$ otherwise. The map

$$\begin{aligned} q : \ & S \to S/F \\ & s \mapsto [s] \end{aligned} \tag{2.1.34}$$

is then a homomorphism. Thus, an equivalence relation F induces a map (2.1.34) from S to its quotient S/F. Conversely, a map $\phi : S \to S'$ defines an equivalence relation by

$$F(s_1, s_2) = 1 \quad :\Leftrightarrow \quad \phi(s_1) = \phi(s_2), \tag{2.1.35}$$

that is, we identify elements of S that have the same image under ϕ. The target set S' thus becomes the quotient S/F.

There always exists the trivial equivalence relation F_0 on S with $F_0(s_1, s_2) = 1$ only if $s_1 = s_2$.

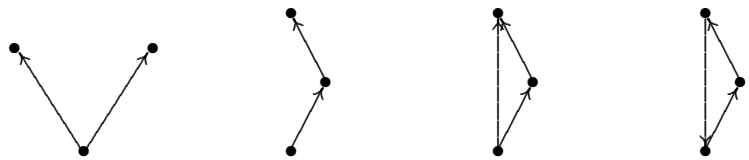
Definition 2.1.5 When the relation F is

reflexive: $F(s, s) = 1$ for all s

transitive: if $F(s_1, s_2) = 1$ and $F(s_2, s_3) = 1$, then also $F(s_1, s_3) = 1$

antisymmetric: if $F(s_1, s_2) = 1$ and $F(s_2, s_1) = 1$, then $s_1 = s_2$,

then we say that (S, F) is a *partially ordered set*, or *poset* for short. One usually writes $s_1 \leq s_2$ in place of $F(s_1, s_2) = 1$ in this case. A partial order provides some (partial) ranking of the elements of S.



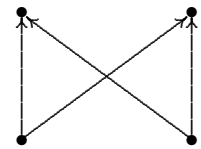
Again, the second graph here is not transitive, nor is the fourth, and therefore they do not represent posets, whereas the other two do. Also, any graph with two arrows in opposite directions between the same two vertices would not be antisymmetric and hence would not represent a poset.

Definition 2.1.6 A *lattice* is a poset $(S, \leq)$ for which any two elements s_1, s_2 have a unique greatest lower bound, that is, there exists some $\underline{s}$, also written as $s_1 \wedge s_2$ and called the *meet* of s_1 and s_2, with

$$\underline{s} \leq s_1, \ \underline{s} \leq s_2, \ \text{and } s \leq \underline{s} \text{ whenever } s \leq s_1, s \leq s_2, \tag{2.1.36}$$

and a unique least upper bound $\bar{s}$, also written as $s_1 \vee s_2$ and called the *join* of s_1 and s_2, with

$$s_1 \leq \bar{s}, \ s_2 \leq \bar{s}, \ \text{and } \bar{s} \leq s \text{ whenever } s_1 \leq s, s_2 \leq s. \tag{2.1.37}$$



This poset is not a lattice as it neither has a unique greatest lower nor a unique least upper bound.

We leave it to the reader to check that the operations $\wedge$ and $\vee$ are associative and commutative. Here, associativity of, for instance, $\wedge$ means that always

$$(s \wedge s') \wedge s'' = s \wedge (s' \wedge s''), \tag{2.1.38}$$

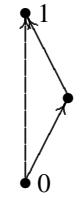
and commutativity that always

$$s \wedge s' = s' \wedge s. \tag{2.1.39}$$

These notions will be taken up in Definitions 2.1.12 and 2.1.13 below.

Definition 2.1.7 We say that the lattice possesses 0 and 1 (not to be confused with the values in (2.1.32)), if it contains elements $0, 1$ with the property that for all $s \in S$

$$0 \leq s \leq 1. \tag{2.1.40}$$



Equivalently, as we leave for the reader to check, a lattice with 0 and 1 is a set with two binary associative and commutative operations $\wedge$ ("and"), and $\vee$ ("or") and two distinguished elements 0, 1, satisfying

$$s \wedge s = s, \ \ s \vee s = s \tag{2.1.41}$$

$$1 \wedge s = s, \ \ 0 \vee s = s \tag{2.1.42}$$

$$s \wedge (s' \vee s) = s \ \ = (s \wedge s') \vee s \tag{2.1.43}$$

for any elements s, s'. The ordering can be recovered from these conditions by stipulating that $s \leq s'$ iff $s \wedge s' = s$, or equivalently, iff $s \vee s' = s'$. The exercise then amounts to verifying that these properties imply that $\leq$ defines an ordering in the sense of Definition 2.1.5, from the properties of $\wedge$ and $\vee$. Thus, here, we can recover a structure from operations; this aspect will be taken up in Sect. 2.1.6 below.

2.1.4.2 Metrics

When the range of F is larger, we obtain more general types of relations. When

$$F : S \times S \to \mathbb{R}, \tag{2.1.44}$$

we obtain the structure of a weighted and directed graph, with $F(s_1, s_2)$ being the weight of the edge from s_1 to s_2.

When we require that

$$F : S \times S \to \mathbb{R}^+ \text{ (the nonnegative real numbers)}, \tag{2.1.45}$$

be symmetric, i.e. $F(s_1, s_2) = F(s_2, s_1)$ for all s_1, s_2, and satisfy the triangle inequality

$$F(s_1, s_3) \leq F(s_1, s_2) + F(s_2, s_3) \text{ for all } s_1, s_2, s_3, \tag{2.1.46}$$

we obtain a *pseudometric*.

When the points s, s' satisfy $F(s, s') = 0$, then, by (2.1.46), also $F(s, \sigma) = F(s', \sigma)$ for all other σ. Therefore, s and s' cannot be distinguished by their relations with other elements in terms of the pseudometric. Therefore, according to the general principle described above, they should be identified. (Above, in the definition of an equivalence relation, we had identified elements with $F(s_1, s_2) = 1$, but, of course, it amounts to the same when we identify elements with $F(s_1, s_2) = 0$ instead. It is an exercise for the reader to check that, for a pseudometric, this does indeed define an equivalence relation.) When we then identify all such equivalent points, we obtain a new set $\bar{S}$, a quotient of the original one, with an induced metric $\bar{F}$. Here, using the standard notation d in place of $\bar{F}$ for a metric, we have

$$d(x_1, x_2) > 0 \text{ whenever } x_1 \neq x_2, \text{ for all } x_1, x_2 \in \bar{S}. \tag{2.1.47}$$

When these conditions are satisfied, $d(.,.)$ is called a *metric*, and we also say that (S, d) is a metric space (the notion of space will be defined below).

A metric provides us with a quantitative notion of nearness, in the sense that we can not only say that for instance y is closer than z to x if

$$d(x, y) < d(x, z), \tag{2.1.48}$$

but we can also quantify that difference.

Examples

1. On the real line $\mathbb{R}$, we have the Euclidean metric

Euclidean metric

$$d(x, y) = |x - y| \text{ for } x, y \in \mathbb{R}. \tag{2.1.49}$$

2. More generally, on $\mathbb{R}^d$, the set of d-tuples $(x^1, \dots, x^d)$, $x^i \in \mathbb{R}$ for $i = 1, \dots, d$, we have the Euclidean metric

$$d(x, y) = \sqrt{\sum_{i=1}^{d} (x^i - y^i)^2} \text{ for } x = (x^1, \dots, x^d),\ y = (y^1, \dots, y^d) \tag{2.1.50}$$

which, of course, reduces to (2.1.49) for $d = 1$.

Trivial metric

3. On any set S, we can define a metric by

$$d(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 = s_2 \\ 1 & \text{if } s_1 \neq s_2. \end{cases} \tag{2.1.51}$$

Thus, any two different points have the same distance from each other. For a set with three points, this looks like

This metric is trivial in the sense that it does not allow any further distinction beyond whether two points are the same or different.

4. A metric d on the set S defines a connected graph if for any $s \neq s' \in S$ there exist $s_0 = s, s_1, \dots, s_n = s' \in S$ with

$$d(s_{i-1}, s_i) = 1 \text{ for } i = 1, \dots, n, \quad \text{and } d(s, s') = n \tag{2.1.52}$$

and we can then let the pair (s_1, s_2) define an edge iff $d(s_1, s_2) = 1$. The first part of the condition then says that any two elements can be connected by a chain of edges. In that sense, the graph is connected. The second part of the condition then specifies that the distance between two vertices of the graph equals the minimal number of edges needed to get from one to the other.

5. On the set of binary strings of some fixed length n, that is, on objects of the form $(b_1 b_2 \dots b_n)$ with $b_i \in \{0, 1\}$, we have the Hamming distance that counts in how many positions two strings $b = (b_1 b_2 \dots b_n)$, $b' = (b'_1 b'_2 \dots b'_n)$ differ, that is,

$$d(b, b') = \sum_{i=1}^{n} |b_i - b'_i|. \tag{2.1.53}$$

6. Whenever $S' \subset S$, a metric on S induces a metric on S' by pullback under the inclusion $i : S' \to S$, see (2.1.2).

When d is a metric on S and $\phi : S' \to S$ a map, then $\phi^* d$ is a metric on S' only if ϕ is injective. Otherwise, it is only a pseudometric, and we need to pass to the quotient where points with the same image under ϕ are identified to obtain a metric, as just described.

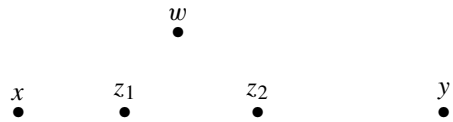
The next definition will be formulated for metric spaces only, although it would also work for pseudometrics. The reason for this restriction is that its content for metric spaces is more interesting and useful than for general pseudometrics.

Definition 2.1.8 Let (S, d) be a metric space. We say that the point $z \in S$ is *between* the points x and y if

$$d(x, y) = d(x, z) + d(y, z), \tag{2.1.54}$$

that is, if the triangle inequality (2.1.46) becomes an equality.

A subset C of S is called *convex* if whenever $x, y \in C$, then also all points that are between x and y are contained in C as well.



The points z_1 and z_2 are between x and y with respect to the Euclidean metric in the plane $\mathbb{R}^2$, but w is not.

For the trivial metric (2.1.51), no point z is between two other points x, y. Consequently, any subset of a set S equipped with that metric is convex.

Definition 2.1.9 Let $p_1, \ldots, p_n \in S$ where (S, d) is a metric space. Then a point $b \in S$ with

$$b = \operatorname{argmin}_q \sum_{i=1,\ldots,n} d^2(p_i, q) \tag{2.1.55}$$

is called a barycenter of $p_1, \ldots, p_n$. Also, a point $m = m(p_1, p_2) \in S$ with

$$d(p_1, m) = d(p_2, m) = \frac{1}{2} d(p_1, p_2) \tag{2.1.56}$$

is called a midpoint of p_1 and p_2.

In particular, a midpoint $m(p_1, p_2)$ is between p_1 and p_2 in the sense of Definition 2.1.8. A midpoint m, if it exists, is also a barycenter of p_1 and p_2. This is easily seen. Let a be any point in S, and let $d(p_1, a) = \lambda d(p_1, p_2)$. (We may assume $\lambda \leq 1$, as otherwise p_2 would yield a lower value in (2.1.55) than a.) By the triangle inequality, $d(p_2, a) \geq (1 - \lambda) d(p_1, p_2)$. Thus,

$$\begin{aligned} d^2(p_1, a) + d^2(p_2, a) &\geq \lambda^2 d^2(p_1, p_2) + (1 - \lambda)^2 d^2(p_1, p_2) \\ &\geq \frac{1}{2} d^2(p_1, p_2) = d^2(p_1, m) + d^2(p_2, m), \end{aligned}$$

and a midpoint thus indeed yields the smallest possible value in (2.1.55).

There is another characterization of midpoints that will become relevant in Sect. 5.3.3. Let

$$B(p, r) := \{q \in S; d(p, q) \leq r\} \text{ for } p \in S, r \geq 0 \tag{2.1.57}$$

be the *closed ball* with center p and radius r. Given $p_1, p_2 \in S$, we then ask for the smallest radius $r = r(p_1, p_2)$ with

$$B(p_1, r) \cap B(p_2, r) \neq \emptyset. \tag{2.1.58}$$

We then observe

Lemma 2.1.3 *The following are equivalent*

(i) $p_1, p_2 \in S$ *possess a midpoint.*
(ii) $r(p_1, p_2) = \frac{1}{2}d(p_1, p_2)$.

When S is finite, barycenters always exist, but midpoints need not. Neither barycenters nor midpoints need be unique. For the metric (2.1.51), there are no midpoints (unless $p_1 = p_2$), but any of the p_i is a barycenter of $p_1, \ldots, p_n$. On a connected graph as characterized by (2.1.52), s and s' possess a (not necessarily unique) midpoint iff their distance is an *even* integer.

2.1.5 Heyting and Boolean Algebras

In this section, we shall consider particular classes of lattices, the Heyting and Boolean algebras. These will play an important role in our discussion of topoi in the last chapter, and they will also arise in our discussion of topologies. Nevertheless, a reader who is primarily interested in the general and abstract aspects of structures may wish to skip this section upon a first reading and only return to it at some later stage.

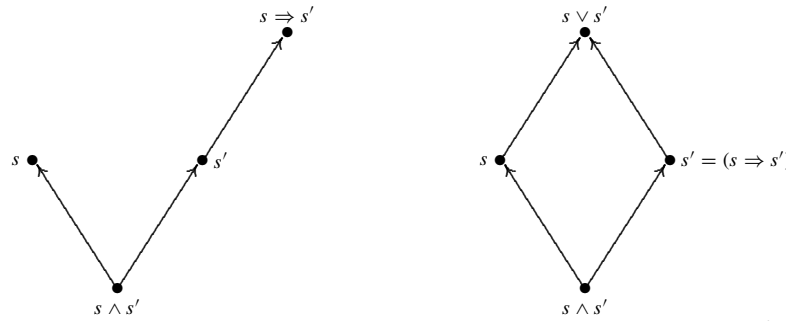
Definition 2.1.10 A lattice with 0 and 1 is a *Heyting algebra* if for any elements s, s', there exists a (unique) element, called the *implication* $s \Rightarrow s'$, satisfying

$$t \leq (s \Rightarrow s') \text{ iff } t \wedge s \leq s'. \tag{2.1.59}$$

The element

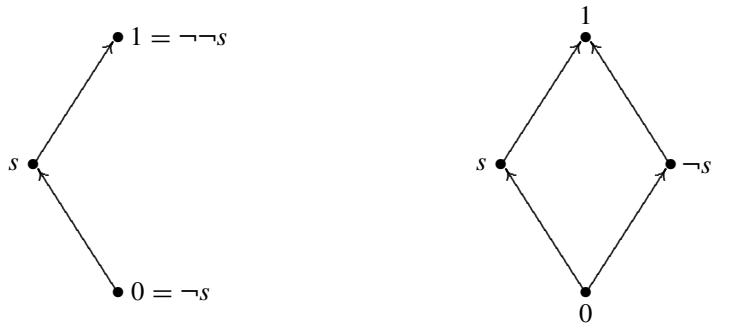
$$\neg s := (s \Rightarrow 0) \tag{2.1.60}$$

is called the *pseudo-complement* of s.



(2.1.61)

In the diagrams (where the arrows from the lowest to the highest vertex that are required by transitivity are not shown), we see that $s \Rightarrow s'$ sits above s', but cannot sit higher above s than s', that is, it still has to satisfy $s \wedge (s \Rightarrow s') = s \wedge s'$, and in fact, it is the highest such element. For the pseudo-complement, we draw the following diagrams (where again some arrows required by transitivity are not shown)



(2.1.62)

In a Heyting algebra, we have

$$(((s \Rightarrow s') \wedge s) \Rightarrow s') = 1, \tag{2.1.63}$$

because $t \leq (((s \Rightarrow s') \wedge s) \Rightarrow s')$ iff $t \wedge (s \Rightarrow s') \wedge s \leq s'$ (by (2.1.59) and the associativity of $\wedge$) iff $t \wedge (s \Rightarrow s') \leq (s \Rightarrow s')$ (by (2.1.59) again), and this is satisfied for all t by the definition of $\wedge$, and finally, any element σ with $t \leq \sigma$ for all t has to be 1, as follows from (2.1.40). In the terminology of elementary logic, (2.1.63) says that the "modus ponens" is always valid. As an exercise for this formalism, you may wish to check that $t \Rightarrow (s \Rightarrow s') = t \wedge s \Rightarrow s'$.

In order to develop the concept of a Heyting algebra further, it is useful to reformulate the definition (2.1.10) of a Heyting algebra without invoking the order relation $\leq$.

Lemma 2.1.4 *A Heyting algebra is a set with two binary associative and commutative operations $\wedge, \vee$ and two distinguished elements $0, 1$, satisfying*

$$x \wedge x = x, \quad x \vee x = x \tag{2.1.64}$$

$$1 \wedge x = x, \quad 0 \vee x = x \tag{2.1.65}$$

$$x \wedge (y \vee x) = x = (x \wedge y) \vee x \tag{2.1.66}$$

for any elements x, y, and a binary operation $\Rightarrow$ characterized as follows. For any elements y, z, there exists a (unique) element $y \Rightarrow z$ satisfying

$$x \wedge (y \Rightarrow z) = x \; iff \; x \wedge y \wedge z = x \wedge y \tag{2.1.67}$$

for all x.

Proof The relations (2.1.64)–(2.1.66) are simply (2.1.41)–(2.1.43), and as explained there, from these operations, we can then recover the order relation $\leq$; it is characterized by $x \leq y$ iff $x \wedge y = x$, or equivalently, iff $x \vee y = y$. □

Therefore, (2.1.67) can also be written as

$$x \leq (y \Rightarrow z) \text{ iff } x \wedge y \leq z \tag{2.1.68}$$

for all x.

From the symmetry of $\wedge$, we get the symmetry

$$x \leq (y \Rightarrow z) \text{ iff } y \leq (x \Rightarrow z). \tag{2.1.69}$$

In the sequel, we leave out the brackets and write, for instance,

$$x \vee y \leq w \Rightarrow z \text{ in place of } (x \vee y) \leq (w \Rightarrow z). \tag{2.1.70}$$

That is, the operations $\vee, \wedge, \Rightarrow$ are carried out before the relation $\leq$ is applied. Similarly for $=$ in place of $\leq$.

Lemma 2.1.5 *A Heyting algebra satisfies the distributive law*

$$(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z) \text{ for all } x, y, z. \tag{2.1.71}$$

Proof For the proof strategy, we observe that

$$x = y \text{ iff for all } w : (x \leq w \text{ iff } y \leq w). \tag{2.1.72}$$

Indeed, this holds in any poset, by inserting $w = x$ and $w = y$ into (2.1.72) and using the antisymmetry of the ordering.

Now $(x \vee y) \wedge z \leq w$ iff $x \vee y \leq z \Rightarrow w$ (by 2.1.68) iff $(x \leq z \Rightarrow w$ and $y \leq z \Rightarrow w)$ iff $(x \wedge z \leq w$ and $y \wedge z \leq w)$ iff $(x \wedge z) \vee (y \wedge z) \leq w$. □

Similarly, we have distributive laws for $\Rightarrow$.

Lemma 2.1.6 *A Heyting algebra satisfies*

$$(x \vee y) \Rightarrow z = (x \Rightarrow z) \wedge (y \Rightarrow z) \tag{2.1.73}$$

and

$$x \Rightarrow (y \wedge z) = (x \Rightarrow y) \wedge (x \Rightarrow z) \tag{2.1.74}$$

for all x, y, z.

Note that we have both $\vee$ and $\wedge$ in (2.1.73), but only $\wedge$ in (2.1.74).

Proof By (2.1.69), $w \leq (x \vee y) \Rightarrow z$ iff $x \vee y \leq w \Rightarrow z$ iff $(x \leq w \Rightarrow z$ and $y \leq w \Rightarrow z)$ iff, using (2.1.69) again, $(w \leq x \Rightarrow z$ and $w \leq y \Rightarrow z)$ iff $w \leq (z \Rightarrow z \wedge y \Rightarrow z)$. This shows (2.1.73). We leave the proof of (2.1.74) to the reader. □

There are some further rules that we shall employ below.

$$x \Rightarrow x = 1 \tag{2.1.75}$$

$$x \wedge (x \Rightarrow y) = x \wedge y \tag{2.1.76}$$

$$y \wedge (x \Rightarrow y) = y \tag{2.1.77}$$

$$(x \Rightarrow (y \wedge x)) = x \Rightarrow y \quad \text{for all } x, y. \tag{2.1.78}$$

For instance, (2.1.78) follows directly from (2.1.71) and (2.1.75), and we leave the other equations as an exercise.

(2.1.60) and (2.1.73) imply one of De Morgan's laws,

$$\neg(x \vee y) = \neg x \wedge \neg y. \tag{2.1.79}$$

(The other De Morgan law, (2.1.84) $\neg(x \wedge y) = \neg x \vee \neg y$, does not hold in all Heyting algebras, but only in Boolean algebras. The reader can check this with the example of $\mathcal{O}(X)$, the algebra of open subsets of a topological space X, introduced in Sect. 4.1 below.)

We conclude our treatment of general Heyting algebras with

Lemma 2.1.7 *When a lattice L with 0 and 1 carries a binary operation $\Rightarrow$ satisfying (2.1.75)-(2.1.78), then this defines a Heyting algebra structure on L.*

Thus, the Heyting algebra structure can be recovered from the two binary operations $\wedge$ and $\Rightarrow$.

Proof We need to show that the two sides of (2.1.67) are equivalent when the conditions (2.1.75)–(2.1.78) hold. Thus, for going from left to right, assume that

$$\begin{aligned} x \wedge (y \Rightarrow z) = x, \qquad & \text{hence} \\ x \wedge y &= x \wedge y \wedge (y \Rightarrow z) \\ &= x \wedge y \wedge z \text{ by (2.1.76)} \end{aligned}$$

so that the right-hand side of (2.1.67) holds. Conversely,

$$\begin{aligned} x &= x \wedge (y \Rightarrow x) \text{ by (2.1.77)} \\ &= x \wedge (y \Rightarrow (y \wedge x)) \text{ by (2.1.78)} \\ &= x \wedge (y \Rightarrow (x \wedge y \wedge z)) \text{ if the r.h.s. of (2.1.67) holds} \\ &= x \wedge ((y \Rightarrow z \wedge y) \wedge (y \Rightarrow x)) \text{ by (2.1.77) (with the roles of } x \text{ and } y \text{ interchanged)} \\ &= x \wedge (y \Rightarrow z) \text{ by (2.1.78) and (2.1.76)} \end{aligned}$$

which is the left-hand side of (2.1.67). □

Definition 2.1.11 Finally, a *Boolean algebra* is a Heyting algebra in which the pseudo-complement $\neg s$ of every element s satisfies

$$s \vee \neg s = 1 \text{ and } s \wedge \neg s = 0. \tag{2.1.80}$$

$\neg s$ is then called the complement of s.

In particular, we then have in a Boolean algebra

$$\neg\neg s = s \text{ for every } s. \tag{2.1.81}$$

In a general Heyting algebra, however, this need not hold; we only have

$$s \leq \neg\neg s. \tag{2.1.82}$$

Thus, the right diagram in (2.1.62) is Boolean, but the left one is not.

Also, the implication and the pseudo-complement in a Boolean algebra are related by

$$s \Rightarrow s' = \neg s \vee s'. \tag{2.1.83}$$

Also, in a Boolean algebra, we have

$$\neg(s_1 \wedge s_2) = \neg s_1 \vee \neg s_2. \tag{2.1.84}$$

Therefore, in a Boolean algebra, the pseudo-complement together with $\vee$ determines $\wedge$ by applying (2.1.81)–(2.1.84)

$$s_1 \wedge s_2 = \neg(\neg s_1 \vee \neg s_2). \tag{2.1.85}$$

Similarly, in a Boolean algebra, $\neg$ and $\wedge$ determine $\vee$,

$$s_1 \vee s_2 = \neg(\neg s_1 \wedge \neg s_2). \tag{2.1.86}$$

A basic example of a Boolean algebra is $\{0, 1\}$ with the above operations, that is,

$$0 \wedge 0 = 0 \wedge 1 = 0, 1 \wedge 1 = 1, 0 \vee 0 = 0, 1 \vee 1 = 0 \vee 1 = 1, \neg 0 = 1. \tag{2.1.87}$$

As a more general example, let X be a set, and let $\mathcal{P}(X)$ be its power set, that is, the set of all its subsets. $\mathcal{P}(X)$ has an algebraic structure with the following operations as $\neg, \vee, \wedge, \Rightarrow$ (for all $A, B \in \mathcal{P}(X)$):

$$\text{Complement:} \quad A \mapsto X \setminus A \tag{2.1.88}$$

$$\text{Union:} \quad (A, B) \mapsto A \cup B \tag{2.1.89}$$

$$\text{Intersection:} \quad (A, B) \mapsto A \cap B := X \setminus (X \setminus A \cup X \setminus B) \tag{2.1.90}$$

$$\text{Implication:} \quad (A, B) \mapsto A \Rightarrow B := (X \setminus A) \cup B. \tag{2.1.91}$$

We note that $C \cap A \subset B$ iff $C \subset (A \Rightarrow B)$, that is, the condition required in (2.1.59).

We also have the relations

$$A \cup (X \setminus A) = X \tag{2.1.92}$$

$$A \cap (X \setminus A) = \emptyset \tag{2.1.93}$$

for all $A \in \mathcal{P}(X)$. Thus, $\emptyset$ and X assume the roles of 0 and 1, resp., that is,

$$\emptyset \subset A \quad (\text{and also } A \cap \emptyset = \emptyset) \tag{2.1.94}$$

and

$$A \subset X \quad (\text{and also } A \cup X = X) \tag{2.1.95}$$

for all $A \in \mathcal{P}(X)$.

Boolean algebra $\{0, 1\} \cong \{\emptyset, X\}$

The Boolean algebra $\{0, 1\} \cong \{\emptyset, X\}$ then arises as the power set of a set X with a single element.

However, when we take X as a set with 2 elements, say $X = \{0, 1\}$,[1] and put

$$\mathcal{O}(X) := \{\emptyset, \{0\}, \{0, 1\}\}, \tag{2.1.96}$$

Heyting algebra $\{\emptyset, \{0\}, \{0, 1\}\}$

then we have a Heyting algebra that is not Boolean (because $\{0\}$ has no complement).

Returning to the case of Boolean algebras, and the above example of $\mathcal{P}(X)$ as an example of such a Boolean algebra, this is in fact an instance of a general relationship as given in the representation theorem of Stone. (As that theorem will not be utilized or referred to in the sequel, and since its proof invokes concepts that may seem a bit technical, some readers may want to skip it on a first reading and continue after (2.1.107).)

[1] Please do not get confused by the different meanings of the symbols 0 and 1 here, as algebraic symbols on one hand and as members of a set on the other hand.

Theorem 2.1.2 (Stone) *For any Boolean algebra B with operations $\vee, \wedge, \neg$, there exist a set X and an injective homomorphism of Boolean algebras*

$$h : B \to \mathcal{P}(X). \tag{2.1.97}$$

Here, a homomorphism $\eta : B_1 \to B_2$ of Boolean algebras has to satisfy $\eta(s_1 \wedge s_2) = \eta(s_1) \wedge \eta(s_2)$ for all $s_1, s_2 \in B_1$ where on the left-hand side, we have the operation $\wedge$ in B_1, and on the right-hand side the operation in B_2, and analogous relations for the other operations $\vee$, $\neg$, and also $\eta(0) = 0$, $\eta(1) = 1$, where again on the left-hand sides, we have the elements 0 and 1 in B_1, whereas on the right-hand sides, the corresponding elements in B_2. (Again, this is an instance of the concept of a homomorphism as a structure preserving map; here, the structure is that of a Boolean algebra, but we have also encountered and will repeatedly encounter homomorphisms for other structure. In Sect. 2.3, the concept will be developed from an abstract perspective.)

Proof A filter $\mathcal{F}$ on B is defined to be a subset of B with the following properties

$$0 \notin \mathcal{F}, \quad 1 \in \mathcal{F}, \tag{2.1.98}$$

$$\text{if } s \in \mathcal{F}, s \leq s', \quad \text{then also } s' \in \mathcal{F}, \tag{2.1.99}$$

$$\text{if } s_1, \ldots, s_n \in \mathcal{F}, \quad \text{then also } s_1 \wedge \cdots \wedge s_n \in \mathcal{F}. \tag{2.1.100}$$

An ultrafilter $\mathcal{F}$ is defined to be a maximal filter, that is, whenever for some filter $\mathcal{G}$

$$\mathcal{F} \subset \mathcal{G}, \text{ then } \mathcal{F} = \mathcal{G}. \tag{2.1.101}$$

Equivalently, $\mathcal{F}$ is an ultrafilter iff

$$\text{for all } s \in B, \text{ either } s \in \mathcal{F} \text{ or } \neg s \in \mathcal{F}. \tag{2.1.102}$$

(Note that as a consequence of (2.1.98) and (2.1.100), s and $\neg s$ cannot both be contained in a filter $\mathcal{F}$.)

The idea of the proof then is to let X be the set of all ultrafilters on B and define

$$\begin{aligned} h : B &\to \mathcal{P}(X) \\ s &\mapsto \{\mathcal{F} : s \in \mathcal{F}\}, \end{aligned} \tag{2.1.103}$$

and verify that this is an injective homomorphism of Boolean algebras. So, let's see how this goes. Constructing filters is easy. For $s \in B$, let $\mathcal{F}(s) := \{s' \in B : s \leq s'\}$. Such an $\mathcal{F}(s)$ is called a principal filter. In general, however, this is not an ultrafilter. We have $\mathcal{F}(s) \subsetneqq \mathcal{F}(\sigma)$ for any $\sigma \leq s, \sigma \neq s$. We can then try to iterate this construction and obtain larger and larger filters that asymptotically yield an ultrafilter. In fact, in general, one has to appeal to the axiom of choice to ensure the existence of ultrafilters.

We observe, however, that by (2.1.98) and (2.1.100), we cannot augment a filter $\mathcal{F}(s)$ so that it contains two elements $s_1, s_2 \leq s$ with $s_1 \wedge s_2 = 0$. Thus, if $s \not\leq t$, then $\neg t \wedge s \neq 0$, and hence we can obtain an ultrafilter containing s, but not t. This yields the injectivity of h.

Among the conditions to check for a homomorphism of Boolean algebras, the most difficult one is

$$h(s_1 \vee s_2) = h(s_1) \cup h(s_2). \tag{2.1.104}$$

We shall check this one and leave the others to the reader. Thus, if $s_1 \vee s_2 \in \mathcal{F}$ for some ultrafilter, then we claim that also s_1 or s_2 is in $\mathcal{F}$. This will then yield $h(s_1 \vee s_2) \subset h(s_1) \cup h(s_2)$. If on the contrary, neither of s_1, s_2 were in $\mathcal{F}$, then by (2.1.102), both $\neg s_1, \neg s_2 \in \mathcal{F}$, hence also $\neg s_1 \wedge \neg s_2 \in \mathcal{F}$ by (2.1.100). But since also, by (2.1.84), $\neg(\neg s_1 \wedge \neg s_2) = s_1 \vee s_2$ is in $\mathcal{F}$ by assumption, this would contradict (2.1.102). For the other direction, that is, $\supset$ in (2.1.104), if $s_1 \vee s_2 \notin \mathcal{F}$, then $\neg s_1 \wedge \neg s_2 = \neg(s_1 \vee s_2) \in \mathcal{F}$, and hence neither s_1 nor s_2 can be in $\mathcal{F}$, using (2.1.98) and (2.1.100) again. □

Actually, an ultrafilter $\mathcal{F}$ on a Boolean algebra B is the same as a homomorphism η from B to the Boolean algebra $\{0, 1\}$; we have

$$\eta(s) = 1 \text{ iff } s \in \mathcal{F}. \tag{2.1.105}$$

We leave it to the reader to check that the properties of ultrafilters ensure that this is indeed a homomorphism of Boolean algebras.

Also, the definition of a filter in the above proof is meaningful for any Heyting algebra H, not necessarily Boolean. Moreover, a filter $\mathcal{G}$ (not necessarily ultra) in H yields a homomorphism $\eta : H \to K$ into another Heyting algebra, with $\eta^{-1}(1) = \mathcal{G}$; this is the natural generalization of (2.1.105). For a principal filter $\mathcal{F}(s)$, this Heyting algebra is simply the poset $H/s := \{s' \in H : s' \leq s\}$ with the Heyting algebra operations induced from H. Equivalently, we can obtain it as the space of equivalence classes for

$$s_1 \equiv s_2 \text{ iff } s_1 \wedge s = s_2 \wedge s. \tag{2.1.106}$$

This construction extends to a general filter $\mathcal{F}$ as

$$s_1 \equiv s_2 \text{ iff } s_1 \wedge s' = s_2 \wedge s' \text{ for some } s' \in \mathcal{F}. \tag{2.1.107}$$

Euclidean space $\mathbb{R}^d$

Let us now consider a structure that is not a Heyting or Boolean algebra. We take the Euclidean space $\mathbb{R}^d$,[2] that is, the space consisting of tuples $x = (x^1, \ldots, x^d)$ with components $x^i \in \mathbb{R}$. The elements of $\mathbb{R}^d$ can be added; with $y = (y^1, \ldots, y^d)$, we have

$$x + y = (x^1 + y^1, \ldots, x^d + y^d) \tag{2.1.108}$$

and multiplied by real numbers α,

$$\alpha x = (\alpha x^1, \ldots, \alpha x^d). \tag{2.1.109}$$

Also, we have the scalar product

$$\langle x, y \rangle = \sum_{i=1}^{d} x^i y^i. \tag{2.1.110}$$

[2] This is, of course, an example of a vector space, but that latter concept will only be introduced below. Therefore, we recall some details here.

A linear subspace of $\mathbb{R}^d$ is a subset of the form

$$V = \{\alpha^1 v_1 + \ldots \alpha^m v_m : \alpha^i \in \mathbb{R}, i = 1, \ldots, d\} \tag{2.1.111}$$

for some elements $v_1, \ldots, v_m \in \mathbb{R}^d$, called generators of V; given V, these v_j are not unique, but this does not matter for our purposes. For two linear subspaces V, W, as is easily checked, their intersection $V \cap W$ is again a linear subspace, and so is their direct sum

$$V \oplus W := \{v + w : v \in V, w \in W\}. \tag{2.1.112}$$

Finally, for a linear subspace, we have the complementary subspace

$$V^\perp := \{w \in \mathbb{R}^d : \langle v, w\rangle = 0 \text{ for all } v \in V\}. \tag{2.1.113}$$

This is where we need the scalar product. Again, the complementary subspace of a linear subspace is indeed itself a linear subspace.

When we then take the intersection $\cap$ of linear subspaces as $\wedge$, the direct sum $\oplus$ as $\vee$ and the complement $^\perp$ as $\neg$, $\mathbb{R}^d$ itself as 1, and the trivial subspace $\{0\}$ (where 0 stands for the element $(0, \ldots, 0)$ of $\mathbb{R}^d$ whose components are all 0) as 0,[3] the linear subspaces of $\mathbb{R}^d$ do not constitute a Heyting algebra. For instance, the distributive law of Lemma 2.1.5 is not satisfied. Some subspace $W \neq \{0\}$ may intersect two other subspaces V_1, V_2 only at 0, but may nevertheless be contained in the direct sum $V_1 \oplus V_2$. For example, consider $\mathbb{R}^2$ spanned by the vectors $e_1 = (1, 0), e_2 = (0, 1)$ and let V_1, V_2, W be the one-dimensional subspaces spanned by $e_1, e_2, e_1 + e_2 = (1, 1)$, resp. Then $V_1 \oplus V_2 = \mathbb{R}^2$ and hence this space contains W while neither V_1 nor V_2 does. In such a case, $(V_1 \oplus V_2) \cap W = W$, whereas $(V_1 \cap W) \oplus (V_2 \cap W) = \{0\}$. For those readers who went through the proof of Theorem 2.1.2, it is also instructive to see why that proof does not apply to this example. In fact, an ultrafilter $\mathcal{F}$ would have to correspond to a smallest subspace $\neq \{0\}$, that is, a subspace W generated by a single vector $w \neq 0$. In other words, the linear filter $\mathcal{F}$ would then consist of all linear subspaces V containing that W. But then (2.1.102) does not hold in general. In fact, there are many subspaces W for which neither W nor $W^\perp$ contain V. For example, take W as above as the span of $(1, 1)$ in $\mathbb{R}^2$, and V as the span of $(1, 0)$. Then neither V nor $V^\perp$, the space spanned by $(0, 1)$, contain W.

In any case, returning to (2.1.32), as soon as we have a basic distinction between two values, 0 and 1 (as in (2.1.32)), false or true, out or in, or whatever, then we can use that distinction to define relations between the elements of a set S.

2.1.6 Operations

Instead of relations, one can also consider *operations*. These are transformations of the elements of a set. When we have a structure, the operations

[3] Carefully distinguish the different meanings of the symbol 0 employed here!

are required to preserve that structure, in a sense defined below. The operations themselves usually constitute a structure. Such a structure can operate on itself, as in the case of a group, or it can operate on another structure, as for the representation of a group. In any case, operations bring us into the realm of *algebra*.

Operations also offer a new perspective on the issue of equivalence. Given an operation of some structure on a set, one can consider two elements of the latter as equivalent when there is an operation moving one of them into the other. In this way, one can form the quotient of the set by the operation, by identifying two elements that are related by an operation, or as one also says in this context, that can be mapped to each other. In order that this be an equivalence relation, so that we can indeed construct such a quotient, we need to require that this operation be reflexive, that is, every element can be mapped to itself (that is, doing nothing counts as an operation), symmetric, that is, when a can be mapped to b, then also b can be mapped to a, and transitive, that is, when a can be mapped to b, and b to c, then we can also map a to c. In terms of the operations, this means that we have an identity operation and that operations can be inverted and composed. When this composition is associative, then the operations constitute a group.

In this sense, we have the principle that any structure should be divided by its automorphism group, the concept of automorphism being defined below.

Let us introduce or recall the basic structures that are defined in terms of operations.

Definition 2.1.12 A *monoid* M is a set each element of which defines an operation

$$\begin{aligned} l_g : \ & M \to M \\ & h \mapsto gh. \end{aligned} \tag{2.1.114}$$

In fact, we shall usually write this as a binary operation

$$(g, h) \to gh \tag{2.1.115}$$

mapping a pair of elements g, h of M to their product gh. This product has to be *associative*

$$(gh)k = g(hk) \text{ for all } g, h, k \in M, \tag{2.1.116}$$

and there must exist a distinguished element e (called the *neutral element*) with

$$eg = ge = g \text{ for all } g \in M. \tag{2.1.117}$$

For instance, a lattice with 0 and 1 possesses two such binary operations, $\wedge$ and $\vee$. According to (2.1.42), the neutral element for $\wedge$ is 1, but the neutral element for $\vee$ is 0.

On the set $\{0, 1\}$, we have two monoid structures, with the operations denoted by $\cdot$ and $+$, resp.,

$$0 \cdot 0 = 0,\ 0 \cdot 1 = 0,\ 1 \cdot 0 = 0,\ 1 \cdot 1 = 1 \text{ and} \tag{2.1.118}$$

$$0 + 0 = 0,\ 0 + 1 = 1,\ 1 + 0 = 1,\ 1 + 1 = 0. \tag{2.1.119}$$

Monoid structures on $\{0, 1\}$

Both structures will be important. Of course, this is also a special case of the observation we have just made about lattices, because we can let $\cdot$ correspond to $\wedge$ and $+$ to $\vee$ in the lattice with 0 and 1 only.

The operation l_g is called the left translation by g. Equivalently, we could write (2.1.115) in terms of the right translation r_h by h. Expressed in terms of such translations, (2.1.117) means that the left and right translations l_e, r_e by the neutral element are the identity operations on M.

Definition 2.1.13 A *group* G is a monoid in which each $g \in G$ has to possess an *inverse* $g^{-1} \in G$ satisfying

$$gg^{-1} = g^{-1}g = e. \tag{2.1.120}$$

The element e and the inverse g^{-1} of a given element g are uniquely determined, as is easily verified.

Definition 2.1.14 A subset S of a group G is called a set of *generators* of the group G if every element of G can be expressed as a product of elements from S and their inverses. (Such a set of generators is not unique.) The group is called *free* if it does not possess nontrivial relations. This means that there exists a set S of generators such that any element of G can be written in a unique way as the product of elements of S and their inverses, apart from inserting trivial products of the form gg^{-1}. (Again, S itself is not unique here.) G is called *torsionfree* if $g^n \neq e$ for all $g \in G, n \in \mathbb{Z}, n \neq 0$.

Free groups are torsionfree, because if $g^n = e$ nontrivially, then e can be expressed in more than one way as a product in G, and hence the same would hold for any other element, e.g., $h = g^n h$.

Definition 2.1.15 The monoid M or the group G is called *commutative* or, equivalently, *abelian* if

$$gh = hg \text{ for all } g, h \in M \text{ or } G. \tag{2.1.121}$$

For a commutative group, the operation is often written as $g + h$, with $-h$ in place of h^{-1}, and the element e is denoted by 0.

Of course, there is the trivial group containing a single element e, with $e \cdot e = e$. The smallest nontrivial group is given by (2.1.119), and we now write this as $\mathbb{Z}_2 := (\{0, 1\}, +)$ with $0+0 = 0 = 1+1,\ 0+1 = 1+0 = 1$. When we consider the same set with a different operation, (2.1.118), which we now write as $M_2 := (\{0, 1\}, \cdot)$ with $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0,\ 1 \cdot 1 = 1$, we obtain a monoid that is not a group (because 0 has no inverse).

Commutative group $\mathbb{Z}_2 := (\{0, 1\}, +)$

Monoid $M_2 := (\{0, 1\}, \cdot)$

More generally, for $q \geq 2$, we can consider the cyclic group $\mathbb{Z}_q := (\{0, 1, \ldots, q-1\}, +)$ with addition defined modulo q, that is, $m + q \equiv m$ for all m. Thus, for instance, $1 + (q-1) = 0$ or $3 + (q-2) = 1$. We can also equip this set with multiplication modulo q, obtaining again a monoid M_q which is not a group.

Cyclic group $(\mathbb{Z}_q, +)$

Monoid $(M_q, \cdot)$

Monoid $\mathbb{N}_0$
Group of integers $\mathbb{Z}$

The nonnegative integers with addition also constitute a monoid $\mathbb{N}_0$; this monoid, however, can be extended to the group $\mathbb{Z}$ of integers.

Group $(\mathbb{Q}_+, \cdot)$

Also, the positive rational numbers $\mathbb{Q}_+$, and likewise the nonzero rationals $\mathbb{Q} \setminus \{0\}$, equipped with multiplication constitute a group.

Definition 2.1.16 A subgroup H of a group G is simply some $H \subset G$ that also forms a group law under the group operation of G. That is, whenever $h, k \in H$, then also $hk \in H$ and $h^{-1} \in H$. Thus, H is closed under the group operation of G, that is, whenever we apply the group operations of multiplication or take the inverse of elements in H, then the result is still in H.

Subgroup $m\mathbb{Z}$ of $\mathbb{Z}$
Subgroups of $\mathbb{Z}_q$

We shall discuss the more abstract Definition 2.3.5 below. Every group G has the trivial group $\{e\}$ and G itself as subgroups. As a nontrivial example, $m\mathbb{Z} := \{\ldots, -2m, -m, 0, m, 2m, \ldots\}$ is a subgroup of $\mathbb{Z}$. Also, when p divides $q \in \mathbb{N}$, then $\{0, p, 2p, \ldots\}$ is a subgroup of $\mathbb{Z}_q$. When q is a prime number (that is, if $q = mn$ with positive integers m, n, then either $m = q, n = 1$ or the other way around), however, then $\mathbb{Z}_q$ does not possess any nontrivial subgroups. Thus, here an arithmetic property, that q be prime, is translated into a group theoretical one, that $\mathbb{Z}_q$ has no nontrivial subgroups.

Ring $(\mathbb{Z}, +, \cdot)$

In fact, the integers $\mathbb{Z}$ also possess another operation, namely multiplication. This leads us to the next

Definition 2.1.17 A *ring* R possesses the structure of a commutative group, written as $+$ (often called addition), and another operation (called multiplication) that is associative, see (2.1.116), and which is *distributive* over $+$,

$$g(h+k) = gh+gk \text{ and } (h+k)g = hg+kg \text{ for all } g, h, k \in G. \quad (2.1.122)$$

The ring is said to be *commutative* when the multiplication is also commutative, see (2.1.121). It is said to possess an *identity* or *unit* if there is an element, denoted by 1, satisfying

$$g1 = 1g = g \text{ for all } g \in R. \quad (2.1.123)$$

Since $0 + 0 = 0$, the distributive law (2.1.122) implies

$$g0 = 0g = 0 \text{ for all } g \in R. \quad (2.1.124)$$

A ring with identity thus possesses a group structure (addition) as well as a monoid structure (multiplication) that is distributive over the group structure.

Ring $(\mathbb{Z}_q, +, \cdot)$

For instance when we equip $\mathbb{Z}_q$ both with addition $+$ and multiplication $\cdot$ modulo q, we obtain a ring. The simplest example is, of course, $\mathbb{Z}_2$ with the two operations given in (2.1.118) and(2.1.119).

More generally, we can also form amalgams of the operations of addition and multiplication.

Definition 2.1.18 A *module* M over a ring R is an abelian group (with group operation denoted by $+$) whose elements can be multiplied by elements of R, denoted by $(r, g) \mapsto rg$ for $r \in R$, $g \in M$, with the following distributive and associative rules.

$$r(g+h) = rg + rh \tag{2.1.125}$$

$$(r+s)g = rg + rs \tag{2.1.126}$$

$$(rs)g = r(sg) \tag{2.1.127}$$

for all $r, s \in R$, $g, h \in M$.

If R possesses an identity 1, then the R-module M is called *unitary* if

$$1g = g \text{ for all } g \in M. \tag{2.1.128}$$

Of course, each ring is a module over itself, as well as a module over any subring[4] of itself. In Definition 2.1.21 below, we shall also consider subsets of a ring that are closed under multiplication of that ring, hence also constitute modules. In those cases, the operation of multiplication is already internal to the structure, but the concept of a module also allows us to consider the multiplication by elements of the ring as something additional, superimposed upon the internal group structure of M. In particular, the elements of R are not considered as elements of M, but rather as operations on M. In the sequel, we shall often encounter such modules over rings. In particular, at several places, we shall construct structures from a ring on which that ring then operates.

In any case, it is an important principle that one structure can operate on another one. In this book, we shall mostly interpret such operations as multiplications, but in other contexts they might arise as translations, time shifts (as an operation by the (positive) real numbers or integers), or whatever. As we shall see below, there are many examples where not a ring, but only a monoid or group operates.

We now come to an important special class of rings.

Definition 2.1.19 A commutative ring R with identity $1 \neq 0$ for which $R \setminus \{0\}$ also is a group under multiplication, i.e., for which every $g \neq 0$ possesses a multiplicative inverse g^{-1} with

$$gg^{-1} = 1, \tag{2.1.129}$$

is called a *field*.

A unitary module over a field is called a *vector space*.

We have already seen an example of a vector space, Euclidean space $\mathbb{R}^d$, see (2.1.108) and (2.1.109).

Euclidean space $\mathbb{R}^d$

[4] In Definition 2.1.16, we have explained what a subgroup is, and you should then easily be able to define a subring, if you do not already know that concept.

Field $\mathbb{Z}_2 = (\{0, 1\}, +, \cdot)$

Vector space $\mathbb{Z}_2^n$

Field $(\mathbb{Z}_p, +, \cdot)$ for prime p

An example of a field is $\mathbb{Z}_2 = (\{0, 1\}, +, \cdot)$ with the operations as defined above.[5] We then have the vector space $\mathbb{Z}_2^n$ over the field $\mathbb{Z}_2$. That vector space consists of all binary strings of length n, like (1011) ($n = 4$) with componentwise addition modulo 2. For instance, $(1100)+(0110) = (1010)$ in $\mathbb{Z}_2^4$. The operation of the field $\mathbb{Z}_2$ on this vector space is given by $0 \cdot a = a, 1 \cdot a = a$ for all $a \in \mathbb{Z}_2^n$. And we have the simple rule that $a + b = 0 \in \mathbb{Z}_2^n$ iff $a = b$.

More generally, $\mathbb{Z}_q$ with the above ring structure is a field if and only if q is a prime number. When q is not prime, there exist elements without multiplicative inverses, namely the divisors of q. For example, in $\mathbb{Z}_4$, we have $2 \cdot 2 = 0 \bmod 4$.

The topic of rings and fields and the relations between them will be taken up in detail in Sect. 5.4.1.

Finally,

Definition 2.1.20 An *algebra* A is a module over a commutative ring R that possesses a bilinear multiplication, that is,

$$\begin{aligned}(r_1a_1 + r_2a_2)b &= r_1a_1b + r_2a_2b \text{ for all } a_1, a_2, b \in A, r_1, r_2 \in R\\ a(r_1b_1 + r_2b_2) &= r_1ab_1 + r_2ab_2 \text{ for all } a, b_1, b_2 \in A, r_1, r_2 \in R\end{aligned} \tag{2.1.130}$$

(Here, for instance, a_1b denotes the multiplication of the two elements of the algebra, whereas r_1a is the multiplication of the element a of A by the element r_1 of the ring R.)

Of course, every ring R is not only a module, but also an algebra over itself. In that case, multiplication in the algebra and multiplication by an element of the ring is the same.

Less trivial, but typical examples are algebras of functions. As those will play an important role later on, let us systematically go through the construction. When U is a set, then the functions from U to a monoid, group, or ring constitute a monoid, group, or ring themselves. (This will be discussed from a more general perspective in Sect. 2.1.7.) For instance, when M is a monoid, and $f : U \to M, g : U \to M$, then for $x \in U$, we can simply put

$$(fg)(x) := f(x)g(x), \tag{2.1.131}$$

as the latter multiplication takes place in the monoid M. Moreover, we can multiply such a function $f : U \to M$ by an element m of M,

$$(mf)(x) := mf(x). \tag{2.1.132}$$

[5] Thus, we have equipped the group $\mathbb{Z}_2$ now with an additional operation, multiplication, but still denote it by the same symbol. We shall, in fact, often adopt the practice of not changing the name of an object when we introduce some additional structure or operation on it. That structure or operation will then henceforth be implicitly understood. This is a convenient, but somewhat sloppy practice. Probably, you will not need to worry about it, but as a mathematician, I should at least point this out.

From this, we see that the functions on U with values in a commutative ring form an algebra. Whether the set U also possesses an algebraic structure is irrelevant here.

Here is another construction of an algebra. Let $\gamma : R \to S$ be a homomorphism of commutative rings. Then S becomes an algebra over R. We have the addition and multiplication in S, and $(r, s) \mapsto \gamma(r)s$ yields the multiplication by elements of R, that is, the module structure for S. The multiplication in S then clearly satisfies the bilinearity laws (2.1.130).

$\mathbb{N}$

Let us also systematically go through the example that you probably know best: The positive integers $\mathbb{N} = \{1, 2, 3, \dots\}$ together with addition do not form a monoid, as the neutral element is missing. This is easily remedied by including 0 and considering the nonnegative integers $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ which form an additive monoid. This monoid is not a group because its elements, except 0, do not have inverses. This is fixed by enlarging it to the additive group of integers $\mathbb{Z}$. If we also include the operation of multiplication of integers, then $\mathbb{Z}$ becomes a ring. This ring is not a field, because apart from 1 and -1, its non-zero elements do not possess multiplicative inverses. Again, we can enlarge it, this time to obtain the field $\mathbb{Q}$ of rational numbers. ($\mathbb{Q}$ can be further enlarged to the fields of the real, the complex, or the p-adic numbers, but this is not our present concern.) All this may appear rather easy in the light of the concepts that we have developed. We should, however, remember that each such extension was an important step in the history of mathematics, and that it provided a crucial motivation for the corresponding abstract concept.

Monoid $\mathbb{N}_0$

Ring $\mathbb{Z}$

Field $\mathbb{Q}$

Definition 2.1.21 A (left) *ideal* I in a monoid M is a subset of M with

$$mi \in I \text{ for all } i \in I, m \in M. \tag{2.1.133}$$

An *ideal* in a commutative ring R with identity is a nonempty subset that forms a subgroup of R as an abelian group and satisfies the analogue of (2.1.133) w.r.t. multiplication.

An ideal in a commutative ring is then also a module over that ring in the sense of Definition 2.1.18.

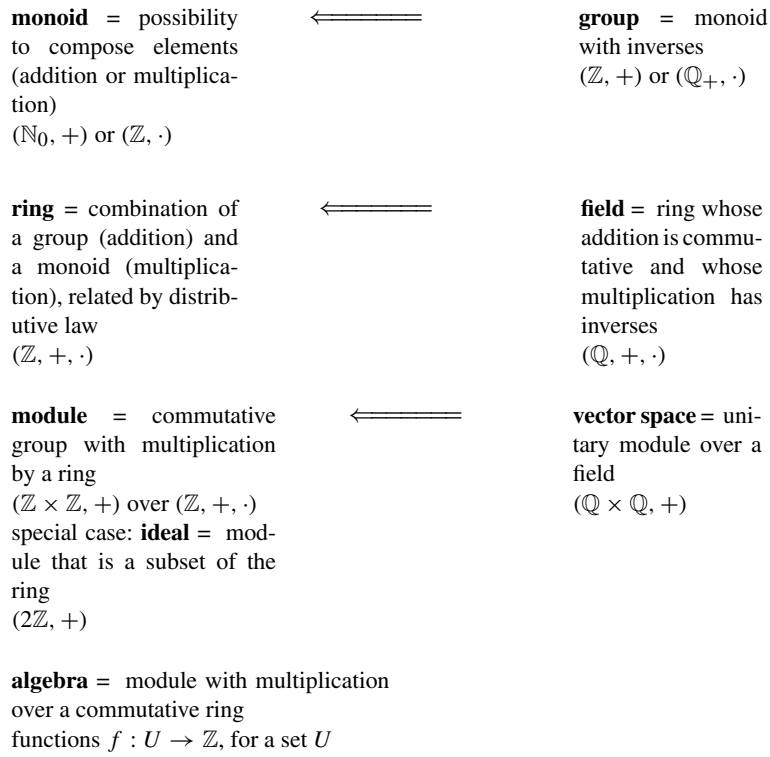
With this concept of an ideal, one can characterize the groups G as precisely those monoids for which $\emptyset$ and G itself are the only ideals. Analogously, a commutative ring R with identity is a field if its only left ideals are $\{0\}$ and R itself. This will be of fundamental importance in Sect. 5.4 below.

When M is the additive monoid $\mathbb{N}_0$ of nonnegative integers, then its ideals are $\emptyset$ and all sets $\{n \geq N\}$ for some fixed $N \in \mathbb{N}_0$. The set of left ideals of the monoid M_2 is $\Lambda_{M_2} := \{\emptyset, \{0\}, \{0, 1\}\}$. We note that this is the same as the above Heyting algebra $\mathcal{O}(X)$ for the 2-element set $X = \{0, 1\}$. More generally, for M_q, the ideals are $\emptyset$, $\{0\}$, M_q and all subsets of the form $\{0, m, 2m, \dots, (n-1)m\}$ when $nm = q$ for $n, m > 1$, that is, nontrivial divisors of q. Thus, when q is a prime number, M_q has only the three trivial ideals, but when q is not prime, there are more. These then are also the ideals of the ring $(\mathbb{Z}_q, +, \cdot)$

Ideals of monoids M_2 and M_q

Ideals of ring $\mathbb{Z}$

The ideals of the ring of integers $\mathbb{Z}$ are of the form $\{nm : n \in \mathbb{Z}\}$ for fixed $m \in \mathbb{Z}$. (Note that while for the monoid $\mathbb{N}_0$, we considered ideals with respect to the operation of addition, for the ring $\mathbb{Z}$, we consider ideals with respect to multiplication.)



The various algebraic structures and the relations between them, with examples. The arrow is an implication, in the sense that, e.g., every group is a monoid; structures build upon those above them.

Symmetric group $\mathfrak{S}_n$

So far, we have looked at the prototypes of abelian groups, $\mathbb{Z}_q$ and $\mathbb{Z}$. In many respects, however, the most important group is the symmetric group $\mathfrak{S}_n$, the group of permutations of n elements, with the composition of permutations as the group operation. We think of these elements as ordered in the form $(1, \ldots n)$. A permutation of them is then written as $(i_1 i_2 \ldots i_n)$, meaning that the element $i_k \in \{1, 2. \ldots, n\}$ is put in the place of k, for $k = 1, 2, \ldots, n$. Of course, the i_j all have to be different, so as to exhaust the set $\{1, 2, \ldots, n\}$. The original ordering $(12 \ldots n)$ then stands for the identity permutation that leaves every element unchanged. The symmetric group $\mathfrak{S}_2$ consists of the two elements (12) and (21) and is therefore isomorphic to $\mathbb{Z}_2$ (we shall explain the term "isomorphic" only below in Definition 2.3.2, but you will probably readily understand its meaning in the present example). The group $\mathfrak{S}_3$ contains 6 elements, (123), (213), (132), (321), (231), (312). (123) is the neutral element. (213), (132) and (321) simply exchange two elements; for instance, (213) exchanges the first and the second element and leaves

$\mathfrak{S}_3$

the third one in place. Each of these three permutations is its own inverse. In contrast, (231) and (312) permute the three elements cyclically, and they are inverses of each other, i.e., $(231) \circ (312) = (123)$. Moreover, $(231) = (132) \circ (213)$, that is, we first exchange the first two elements (note the reversal of order in our notation; here, the permutation (213) is carried out first, and then the permutation (132) is applied), and then the last two. After the first exchange, 1 is in the middle position, and the second exchange then moves it into the last position. Also, $(231) = (213) \circ (321)$, whereas $(213) \circ (132) = (312)$. In particular, $\mathfrak{S}_3$ is not abelian, since $(132) \circ (213) \neq (213) \circ (132)$.

In order to simplify the notation, one can leave out those elements that are not affected by the permutation. Thus, instead of (132), one simply writes (32) for the exchange of the second and the third element. With this notation, for example, $(21) \circ (32) = (312)$. Again, note the reverse order of the operation: We first exchange the last two elements, which brings 3 into the second position, and we then exchange the first two positions, which then brings 3 from the second into the first position.

We can also already make some general observations. The group $\mathfrak{S}_m$ is contained in $\mathfrak{S}_n$ for $m < n$—just take m of the n elements and permute them and leave the remaining $n - m$ elements alone. $\mathfrak{S}_m$ is a subgroup of $\mathfrak{S}_n$ in the terminology of Definition 2.1.16 or 2.3.5 below. This means that the inclusion that we have described defines $i : \mathfrak{S}_m \to \mathfrak{S}_\mathfrak{n}$ such that for all $g, h \in \mathfrak{S}_m$, we have

$\mathfrak{S}_n$

$$i(g \circ h) = i(g) \circ i(h) \tag{2.1.134}$$

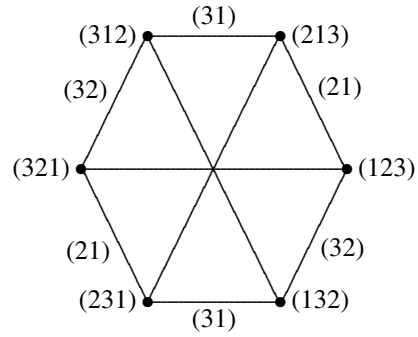
where the first $\circ$ is the multiplication in $\mathfrak{S}_m$ and the second one that in $\mathfrak{S}_n$. More generally, a map $i : G_1 \to G_2$ between groups or monoids satisfying (2.1.134) is called group or monoid homomorphism. Thus, a homomorphism is a map compatible with the group or monoid operations. Analogously, one may then define homomorphisms of rings or fields.

Returning to the example of the symmetric groups $\mathfrak{S}_n$, we see that, since $\mathfrak{S}_3$ is not abelian, this implies that neither are the groups $\mathfrak{S}_n$ for $n > 3$.

Also, when G is a group with finitely many elements, or a finite group for short, then left multiplication l_g by any $g \in G$ induces a permutation of the elements of G. For this, we simply observe that $l_g : G \to G$ is injective, because if $gh = gk$, then also $h = g^{-1}(gh) = g^{-1}(gk) = k$. That is, l_g maps different elements to different elements, and it therefore permutes the elements of g. Likewise, the assignment $g \to l_g$ is injective, in the sense that if $g_1 \neq g_2$, then also $l_{g_1} \neq l_{g_2}$, as for instance $l_{g_1} e = g_1 \neq g_2 = l_{g_2} e$.

A group G defines a graph. More precisely, take a group G with a set S of generators that is closed under inversion, that is, whenever $g \in S$, then also $g^{-1} \in S$ (one might simply start with any set S' of generators and enlarge S' to S by all the inverses of its elements). The so-called Cayley graph of the pair (G, S) then has as vertex set the elements of G, and there is an edge between $h, k \in G$ iff there is some $g \in S$ with $gh = k$. Since then also g^{-1} by our condition on S and $g^{-1}k = h$, this relation between h and k is symmetric, and the graph is therefore undirected. For instance, for the symmetric group $\mathfrak{S}_3$, we can take the generators (21), (32), (31) each of which is its own inverse. The resulting Cayley graph is then

Cayley graph of $\mathfrak{S}_3$



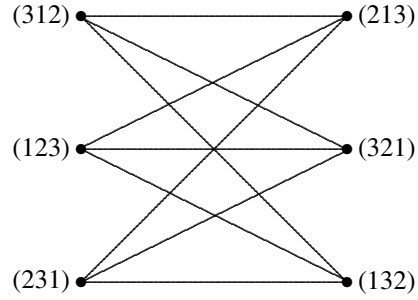
(2.1.135)

In this graph, in order to avoid clutter, we have labelled only some of the edges, and for the edges, we have used the abbreviated notation explained above. We also see that this graph is bipartite, that is, there are two classes of nodes, {(123), (231), (312)} and {(132), (321), (213)} with the property that there are only edges between nodes from different classes. In fact, the second class consists of permutations of two elements, in our shorthand notation {(32), (31), (21)}. Such permutations of two elements are called transpositions. The elements of the other class are products of an even number of such transpositions. In fact, each symmetric group $\mathfrak{S}_n$ contains the two subclasses of elements that can be written as products of an even or of an odd number of transpositions. (We can also formulate this slightly more abstractly and define the parity or sign $\operatorname{sgn}(g)$ of an element g of $\mathfrak{S}_n$ to be 1 (even) if it can be expressed as even number, and -1 (odd) if it can be expressed as an odd number of transpositions. Of course, we then have to verify that the parity is well defined, in the sense that there is no permutation that can be represented by both an even and an odd number of transpositions. The essential fact underlying this is that the product of two transpositions—which is even—can never be a single transposition itself—which would be odd. This, however, is readily checked. Also, $\operatorname{sgn}(gh) = \operatorname{sgn}(g)\operatorname{sgn}(h)$. In more abstract terminology, this means that sgn is a group homomorphism from $\mathfrak{S}_n$ to the group $\{1, -1\}$ with multiplication as the group law. In turn, the assignment $1 \to 0, -1 \to 1$ is a homomorphism of this group to the group $\mathbb{Z}_2$ with addition as the group law.)

The first subclass, the even products, actually forms a subgroup of $\mathfrak{S}_n$. This group is called the alternating group $\mathfrak{A}_n$.

Cayley graph of $\mathfrak{S}_3$

In order to show the bipartiteness of this Cayley graph, we rearrange the positions of the nodes to get the following figure.

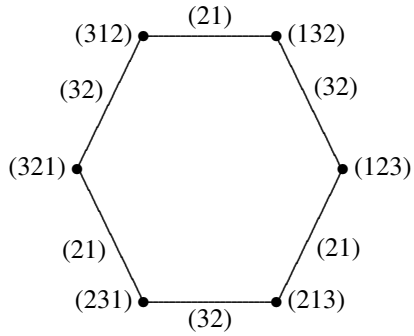


(2.1.136)

In fact, we have here a *complete* bipartite graph, meaning that every node in one class is connected with every node in the other class. Let us emphasize that (2.1.136) depicts the same graph as (2.1.135). Rearranging the positions of the nodes can create visualizations that look very different although they represent the same structure.

Here is another observation. The permutation (31) can also be expressed as the product $(31) = (21)(32)(21)$ of three transpositions of *adjacent* elements. Again, this is a general phenomenon. Any element of $\mathfrak{S}_n$ can be written as a product of transpositions of adjacent elements. This product representation is not unique (for instance, also $(31) = (32)(21)(32)$), but the parity (odd vs. even) is invariant. In particular, we can choose the set of transpositions of adjacent elements as a set of generators of $\mathfrak{S}_n$. This would change the above Cayley graph (2.1.135) to the following one

Cayley graph of $\mathfrak{S}_3$



(2.1.137)

From this graph, for instance, we can directly read off the identity $(21)(32)(21) = (32)(21)(32)$.

Of course, we could also take $S = G\backslash\{e\}$, that is, let our set of generators consist of all nontrivial elements of G. The resulting Cayley graph would then be a complete graph, that is, a graph where each element is connected with every other element. In fact, the same is true for any finite group G, i.e., when we take all nontrivial elements of G as generators, the resulting Cayley graph is complete.

The symmetric groups will appear again in Sect. 3.4.4 below. In fact, there we shall see a converse to the preceding observation that we can obtain complete graphs as Cayley graphs of the symmetric (or any other) group: The automorphism group of a complete graph is a symmetric group.

Given such algebraic structures, you can of course combine them to form other such structures. For instance, we can form the product GH of two groups GH. The elements of GH are pairs (g, h), $g \in G, h \in H$, and the group operation is likewise assembled from those in G and H,

$$(g_1, h_1)(g_2, h_2) := (g_1 g_2, h_1 h_2). \tag{2.1.138}$$

Subsequently, we shall utilize another important concept from group theory, that of a normal subgroup. To prepare for this concept, let us consider a group G with a subgroup N. We then consider the set G/N of the equivalence classes for the equivalence relation

$$g \sim h \text{ iff there exists some } n \in N \text{ with } h = ng. \tag{2.1.139}$$

(2.1.139) defines an equivalence relation, indeed, because N is a subgroup, and not just a subset of G. For instance, for the transitivity of (2.1.139), let $g \sim h$ and $h \sim k$, that is, there exist $n, m \in N$ with $h = ng, k = mh$. But then $k = mng$ and since $mn \in N$ because N is a group, we obtain $g \sim k$. Nevertheless, it seems that this leads us outside the realm of groups, as so far, G/N is only a set, although we had started with groups G and N. In other words, we are led to the question of whether G/N can also be turned into a group. It turns out, however, that we need to impose an additional condition upon N for that purpose.

Definition 2.1.22 A subgroup N of a group G is called a *normal* subgroup if for every $g \in G$

$$g^{-1}Ng = N. \tag{2.1.140}$$

Lemma 2.1.8 *If N is a normal subgroup of G, then the quotient G/N can be equipped with a group structure.*

Proof We need to verify that the group multiplication in G passes to the quotient G/N in the sense that we can carry it out unambiguously on equivalence classes. That is, when $[g]$ denotes the equivalence class of $g \in G$, we need to verify that

$$[g][h] := [gh] \tag{2.1.141}$$

is well defined in the sense that it does not depend on the choice of elements in the equivalence classes. That means that when $g' \sim g, h' \sim h$, then we want

$$g'h' \sim gh, \tag{2.1.142}$$

that is,

$$[g'h'] = [gh]. \tag{2.1.143}$$

Now, when $g' \sim g, h' \sim h$, there exist $m, n \in N$ with $g' = mg, h' = nh$. Then

$$g'h' = mgnh = mgng^{-1}gh. \tag{2.1.144}$$

Now since $gNg^{-1} = N$ as N is a normal subgroup, the element gng^{-1} is again an element of N. Calling this element n', we obtain from (2.1.144)

$$g'h' = mn'gh, \tag{2.1.145}$$

and since $mn' \in N$ as N is a group, we conclude that (2.1.143) holds. □

Putting it differently, the Lemma says that we have a group homomorphism

$$\iota : G \to G/N, \quad g \to [g]. \tag{2.1.146}$$

Let us consider some *examples* of normal subgroups.

- Any subgroup of an abelian group is normal, since $g^{-1}ng = n$ in that case.

- When $q = mn$ is not prime, that is, m, n, q are all integers > 1, then $\mathbb{Z}_m$ and $\mathbb{Z}_n$ are subgroups of $\mathbb{Z}_q$ ($\mathbb{Z}_m$ becomes the subgroup with elements $0, m, 2m, \ldots, (n-1)m$ of $\mathbb{Z}_q$, and analogously for $\mathbb{Z}_n$), and since $\mathbb{Z}_q$ is abelian, they are normal subgroups.
- The alternating group $\mathfrak{A}_n$ is a normal subgroup of the symmetric group $\mathfrak{S}_n$, because the parity of $g^{-1}ng$ is the same as the parity of n. In fact, since we have observed above that sgn induces a group homomorphism $\mathfrak{S}_n \to \mathbb{Z}_2$, this also follows from the next example.
- If $\rho : H \to G$ is a group homomorphism, then $\ker(\rho) := \{k \in H : \rho(k) = e \in G\}$ is a normal subgroup. In fact, if $k \in \ker \rho$, then for any $h \in H$, we have $\rho(h^{-1}kh) = \rho(h)^{-1}\rho(k)\rho(h) = \rho(h)^{-1}e\rho(h) = e$, and hence also $\rho(h)^{-1}\rho(k)\rho(h) \in \ker \rho$.

Normal subgroups of $\mathbb{Z}_q$

Alternating group $\mathfrak{A}_n$

Actually, the last example is not really an example, but a general fact. Every normal subgroup N of a group G is the kernel of a homomorphism, the homomorphism $\iota : G \to G/N$ of (2.1.146).

Definition 2.1.23 A group G is called *simple* if it is nontrivial and its only normal subgroups are the trivial group and G itself.

This is a very important concept, because simple groups are the basic building blocks of group theory. A group G that is not simple possesses a nontrivial normal subgroup N and therefore can be broken up into two smaller groups, the normal subgroup N and the quotient group G/N. When one or both of those are still not simple, the process can be repeated. When G is finite, the process then has to terminate after finitely many steps, and one has assembled the building blocks of G (by the Jordan-Hölder theorem, which we do not prove here, they are, in fact, unique). Thus, when one has a list of the simple finite groups, one can then construct all finite groups, and the subject of finite groups is mathematically completely understood. The complete classification of finite simple groups, however, turned out to be quite difficult and was only successfully completed around 2004.

The most important examples of finite simple groups are the cyclic groups $\mathbb{Z}_p$ for a prime p (as we have just seen in the examples of normal subgroups, $\mathbb{Z}_n$ is not simple when n is not prime) and the alternating groups $\mathfrak{A}_n$ for $n \geq 5$ (the latter fact is not completely trivial). There are such infinite series of finite simple groups, and 26 exceptional groups, among which the so-called monster group is the most complex. A recent textbook is [118]. The approach to the classification of finite groups as just described embodies an important principle of mathematical research. Whenever one wants to understand some class of mathematical structures, one first needs to identify the elementary building blocks that are no longer decomposable by some basic operation (here, taking the quotient by a normal subgroup), and then one has to classify those building blocks. All the structures in the class under consideration can then be composed out of those building blocks.

2.1.7 Parametrizing Constructions

We have defined or characterized algebraic structures like Boolean and Heyting algebras, monoids, groups, rings and fields in terms of operations. There were nullary operations—the distinguished elements 0, 1—, unary operations—the pseudo-complement $v \to \neg v$ or the inverse $g \to g^{-1}$—and binary operations—$(v, w) \to v \wedge w$ or $(g, h) \to gh$. Let S be such a structure, and let X be a set. Then, typically, the set S^X of maps

$$\phi : X \to S \tag{2.1.147}$$

again carries the same type of structure. For instance, when S is a group, S^X is also a group, with group law

$$\begin{aligned} (\phi, \psi) \ &\to \phi\psi \\ \text{with} \ \ \phi\psi(x) &:= \phi(x)\psi(x) \end{aligned} \tag{2.1.148}$$

where on the right hand side, we have the group operation of S applied to the elements $\phi(x)$ and $\psi(x)$. We may consider S^X as a family of groups parametrized by the elements of X, but the preceding construction tells us that this family is a group in its own right. So, if we want, we can iterate the construction and consider maps from some other set into S^X. Again, the result will be a group.

The same works for monoids or rings. It does not work, however, for fields. That is, the maps from a set X into a field F do not constitute a field; they only yield a ring (except for the trivial case when X has only one element). The reason is the exceptional role of the distinguished element 0 as being the only one without a multiplicative inverse. In F^X any ϕ that maps some, but not all elements of X to 0 then fails to have a multiplicative inverse, without being the neutral element for addition in F^X—the latter one being the map that maps *all* $x \in X$ to 0.

When we have a relation like in (2.1.32), $F : S \times S \to \{0, 1\}$, we have two immediate possibilities for defining such a relation on S^X,

$$F(\phi, \psi) := \sup_{x \in X} F(\phi(x), \psi(x)) \tag{2.1.149}$$

or

$$F(\phi, \psi) := \inf_{x \in X} F(\phi(x), \psi(x)) \tag{2.1.150}$$

In the first case, we have $F(\phi, \psi) = 1$ if there exists some $x \in X$ with $F(\phi(x), \psi(x)) = 1$. In the second case, this would have to hold for all $x \in X$.

We can do the same when the relation takes values in $\mathbb{R}$ or $\mathbb{R}^+$, provided the supremum or infimum stays finite.

Just as an aside, to put this into the perspective of analysis: If X carries a measure μ (see Sect. 4.4), we can also average this construction w.r.t. μ. For instance, when the relation is a metric $d(.,.)$, and if $1 \leq p < \infty$, we obtain a metric on S^X via

$$d_p(\phi, \psi) := \left(\int_X d^p(\phi(x), \psi(x))\mu(dx)\right)^{1/p} \tag{2.1.151}$$

where $d^p(y, z)$ means $d(y, z)^p$; again, we need to assume that this expression is finite, of course. The case $p = 2$ is the most important one in practice.

The restriction $p \geq 1$ is needed for the triangle inequality to hold for the metric d_p on S^X. d_p is also called the L^p-metric on S^X. The analogue of (2.1.149) is

$$d_\infty(\phi, \psi) := \text{ess sup}_{x \in X} d(\phi(x), \psi(x)), \tag{2.1.152}$$

where ess sup stands for the essential supremum (i.e., ess $\sup_{x \in X} f(x) := \inf\{a \in \mathbb{R} \cup \{\infty\} : f(x) \leq a \text{ for all } x \in X \setminus A\}$ for some $A \subset X$ with $\mu(A) = 0$, a so-called nullset, see Sect. 4.4). (For details, see e.g. [59]). Of course, when X is, for instance, finite, we may replace the essential supremum by the ordinary supremum.

2.1.8 Discrete Versus Continuous

Besides the distinction between geometry and algebra, another important distinction in mathematics is that between discrete and continuous structures. In abstract terms, this is the difference between *counting* and *measuring*. In its basic form, counting utilizes the positive integers $\mathbb{N}$, that is, the possibility of enumeration. Measuring depends on the real numbers $\mathbb{R}$, and the ancient Greeks have already discovered that measuring cannot be reduced to counting, because of the existence of irrational numbers, that is, numbers that cannot be expressed as a ratio of two integers. Real numbers can be expressed as sequences of integers, in fact, even as binary strings, but Cantor's diagonal argument displayed in Sect. 2.1.3 showed us that the reals cannot be enumerated by the integers.

Continuous structures constitute the realm of *analysis*. In analysis, fundamental concepts are limits and convergence, completion, and compactness.

The completely new aspect that analysis offers is that relations need not hold exactly, but only approximatively, and that this can be quantified in the sense that the approximation error can be controlled or estimated. For that purpose, analysis is based on continuous structures which enable us to define perturbations and variations. While such structures constitute the basis of analysis, the methods can then even be applied to discrete settings, as in numerical analysis, where continuous structures are utilized conceptually to interpolate between discrete values, and to prescribe the resolution of a discrete scheme by comparison with an idealized continuous one.

Returning to the above example of a pseudometric F, let us assume that there is also some metric d given on S. Then $F + \epsilon d$, for any positive real number ϵ, yields a perturbation of the pseudometric F that is positive definite, i.e., a metric on S. We may apply this even to the trivial case where $F \equiv 0$. In that case, the above quotient $\bar{S}$ obtained by identifying points s, s' with $F(s, s') = 0$ consists of one single point, whereas the quotient by $F + \epsilon d$ for any $\epsilon > 0$ is the entire set S. Therefore, the quotient construction in this case does not depend continuously on ϵ. In a certain sense, then the algebraic quotient construction and the analytic limit construction $\epsilon \to 0$ are not compatible with each other.

Some mathematical concepts connect two of the above domains, geometry, algebra, and analysis. The concept of a Lie group, for instance, connects all three of them, but this will not be treated in the present book.

2.2 Axiomatic Set Theory

This section will not be referred to in the sequel, and it can therefore be skipped. Its purpose is to provide a formal substrate for some notions that have been employed in the preceding in an intuitive sense. More precisely, we shall briefly describe the Zermelo-Frankel version of axiomatic set theory that started with [123]. There are many textbooks on axiomatic set theory, for instance [109].

The basic notions of a set and of membership in a set (expressed by the symbol $\in$) are not defined in axiomatic set theory, but they are assumed to exhibit certain properties that are stipulated in the axioms. Such axioms should be consistent (in the sense that it is not possible to derive from them both a statement and its negation), plausible, and rich enough to derive for instance the basic results of Cantor's theory of sets.

We now list 10 axioms of set theory that emerged from the work of Zermelo, Frankel, Skolem, von Neumann and others.

1. **Axiom of extension**: *Let A, B be sets. If for all x, $x \in A$ if and only if $x \in B$, then $A = B$.*
   Thus, sets with the same elements are equal. This axiom says that a set is determined by its elements.
2. **Axiom of the empty set**: *There exists a set $\emptyset$ with the property that for all x, $x \notin \emptyset$.*
   Here, $x \notin$ abbreviates "it is false that $x \in$". The empty set thus is simply the set without elements.
3. **Axiom of separation**: *Let A be a set. Then for every definite condition $P(x)$, there exists a set B such that for all x, $x \in B$ if and only* if x satisfies $P(x)$.
   Here, a definite condition is built from the so-called atomic formulas $x \in y$ and $x = y$ (where x and y are variables) through finitely many applications to formulas P, Q of connectives (*if P, then Q; P iff Q; P and Q; P or Q; not P*) and quantifiers (*for all x, P holds (abbreviated as $P(x)$); for some x, Q holds*). In a condition $P(x)$, the variable x must be free, that is, not under the scope of a quantifier. In contrast, for a condition as in the axiom, B must not be free, that is, it has be bound by a quantifier. The concepts of propositional and predicate logic that are employed will be taken up in Sect. 9.3. We have discussed this as a principle for specifying subsets of a given set already in Sect. 2.1.3.
4. **Axiom of pairing**: *If A, B are sets, there exists a set (A, B) that has A and B as its only elements.*
   (A, B) is called the unordered pair of A and B. This and the next two axioms ensure that the application of standard operations to sets produces sets again.
5. **Axiom of union**: *Let A be a set. Then there exists a set C with the property that $x \in C$* iff *$x \in a$ for some $a \in A$.*
   This means that the union of all the sets that are elements of A is again a set. We also write $\bigcup A$ for this set. A more explicit notation would be $\bigcup_{a \in A} a$.

6. **Axiom of power set**: *Let A be a set. Then there exists a set $\mathcal{P}(A)$, called the power set of A, with the property that $B \in \mathcal{P}(A)$ whenever for all x, $x \in B$* implies *$x \in A$.*
   The power set $\mathcal{P}(A)$ thus contains all the subsets of A as its elements. We have discussed the power set already in Sect. 2.1.3 where we have connected the axioms of separation and power set.
7. **Axiom of infinity**: *There exists a set N with the properties that $\emptyset \in N$ and whenever $x \in N$, then also $x \cup \{x\} \in N$.*
   Here, $\{x\}$ is the set having x as its only member, and $x \cup \{x\}$ means that we add to the set x x itself as a further member. The axiom of infinity can be seen as an abstract version of Peano's principle of induction that generates the natural numbers (positive integers). We consider $x \cup \{x\}$ as the successor of x. Such a set N might be written in an iterative fashion as
   $$N = \{\emptyset, \emptyset \cup \{\emptyset\}, \emptyset \cup \{\emptyset\} \cup \{\emptyset \cup \{\emptyset\}\}, \dots\}. \tag{2.2.153}$$
   In fact, in place of $\emptyset$, we could have started with an arbitrary element x. Therefore, more succinctly, we introduce a symbol 1 and write $1' := 1 \cup \{1\}$. Then we obtain such a set N as
   $$N = \{1, 1', 1'', 1''', \dots\}. \tag{2.2.154}$$
8. **Axiom of choice**: *Let A be a set whose elements are nonempty sets. Then there exists a mapping $f : A \to \bigcup A$ with $f(a) \in a$ for all $a \in A$.*
   Thus, for every $a \in A$, we can choose some element $f(a)$ of a.
9. **Axiom of replacement**: *Let A be a set and f a mapping defined on A. Then there exists a set B whose elements are precisely the $f(x)$ for $x \in A$.*
   Thus, the image of a set under a mapping is again a set. As an application, the map $1 \mapsto N$ from (2.2.154) produces the set
   $$\{N, N', N'', \dots\}. \tag{2.2.155}$$
10. **Axiom of restriction**: *Every set A contains an element a* with $A \cap a = \emptyset$.
    Thus, A and its element a have no element in common. This last axiom is only introduced in order to rule out certain undesired models of the first nine axioms. More precisely, it serves to rule out infinite descending chains, that is, $\dots a_n \in a_{n-1} \in \cdots \in a_0$. In particular, according to this axiom, we must not have $a \in a$.

The preceding axioms are not all independent of each other. In fact, several of them can be left out without loss of scope, as they can be derived from the remaining ones. For instance, the axiom of the empty set can be omitted, and so can the axiom of separation which can be deduced from the axiom of replacement. Also, from the latter together with the axiom of power set one can deduce the axiom of pairing. Thus, the list of axioms reflects the historical development rather than their logical status. Also, some people do not accept the axiom of choice.

There is an alternative system of axioms, named after Bernays and Gödel; in some treatises, von Neumann is also included.

In the sequel, we shall assume that we have some fixed universe U of sets that satisfies the above axioms. Any set to be mentioned in the sequel will be assumed to be a member of U.

2.3 Categories and Morphisms

The concepts of *category* and *morphism* unify some (but not all) of the preceding.

Definition 2.3.1 A *category* $\mathbf{C}$ consists of *objects* $A, B, C, \ldots$ and *arrows* or *morphisms*

$$f : A \to B \tag{2.3.1}$$

between objects, called the *domain* $A = \mathrm{dom}(f)$ and *codomain* $B = \mathrm{cod}(f)$ of f. Arrows can be composed, that is, given $f : A \to B$ and $g : B \to C$, there is an arrow

$$g \circ f : A \to C. \tag{2.3.2}$$

(The requirement for the composition is solely that $\mathrm{cod}(f) = \mathrm{dom}(g)$.) This composition is *associative*, that is,

$$h \circ (g \circ f) = (h \circ g) \circ f \tag{2.3.3}$$

for $f : A \to B, g : B \to C, h : C \to D$.

For each object A, we have the *identity arrow* ("doing nothing")

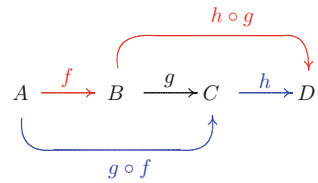
$$1_A : A \to A \tag{2.3.4}$$

which satisfies

$$f \circ 1_A = f = 1_B \circ f \tag{2.3.5}$$

for all $f : A \to B$.

The associativity condition (2.3.3) can also be expressed by saying that whichever sequence of arrows we follow from A to D in the diagram below, the result is the same.



(2.3.6)

We can either follow the red or the blue sequence or the middle one, $h \circ g \circ f$.

Somewhat sloppily, we shall occasionally write $C \in \mathbf{C}$ to say that C is an object of the category $\mathbf{C}$.

One might object here that Definition 2.3.1 is not really a mathematical definition because it is left undefined what an "object" or a "morphism" is

or should be. Whenever we speak of a category, we therefore first need to specify what its objects and morphisms are. For the abstract language of categories, however, it is irrelevant what the objects and morphisms are. They only need to satisfy the rules laid down in Definition 2.3.1.

The idea is that the objects of a category share some kind of structure, and that the morphisms then have to preserve that structure. A category thus consists of objects with structure and directed relations between them. A very useful aspect is that these relations can be considered as operations.

Taking objects as vertices and arrows as edges, we can thus consider a category as a directed graph, with the property that each vertex stands in relation to itself, that is, has an edge from itself to itself. This graph might have multiple edges, as there could be more than one morphism between two objects.

In this sense, the arrows of a category are viewed as relations. One can also view them as operations, as mappings between the objects. An arrow from A to B thus maps A to B.

Viewing morphisms as operations may remind you of the notion of a group, but in contrast to what was required in the Definition 2.1.13 of a group for the left translation l_g by a group element, for the morphisms of the category we do not require that they can be inverted, nor that we can compose any two of them. Nevertheless, we have

Lemma 2.3.1 *A category with a single object is a monoid, and conversely.*

Proof Let M be a monoid, as in Definition 2.1.12. We consider the elements g of M as operations, $h \mapsto gh$, that is, as arrows

$$l_g : M \to M. \tag{2.3.7}$$

Since they have to satisfy the associativity law, they define the morphisms of the category with the single object M. The neutral element e yields the identity morphism 1_M.

Conversely, the arrows of a category with a single object M can be considered as the left translations l_g of M, hence as elements of a monoid, as they satisfy the associativity law. The identity arrow 1_M yields the neutral element e of that monoid. □

Categories can be constructed and considered at different levels of abstraction, as we shall explore in the following. As a brief guide to things that will be explained in more detail shortly, let us point out the following principle: On one hand, the structures that we have considered in the preceding constitute categories. A set, a graph or digraph, a poset, lattice, Heyting or Boolean algebra, a monoid, group, ring, or field are all categories, with the objects being the elements of that structure, and the morphisms being given by the relations or operations within that structure.[6] On the other hand, however, at the next level, the ensemble of structures of a given type also constitute a category. Thus we shall have the category of sets, the category of

[6]Alternatively, as noted above, for an algebraic structure like a group, we could consider that structure as the single object, and its elements as the morphisms of that object.

posets, those of graphs and digraphs, of metric spaces, of monoids, groups, rings, or fields, etc. The morphisms then are structure preserving mappings between two such structures, e.g., between two groups. Thus, within the context of the corresponding category, we can consider all structures of a given type simultaneously and consider the structure preserving relations between them. We can then move to still higher levels of abstraction and consider categories of categories of categories and figure out what the morphisms in that case should be. Or we can consider categories of morphisms. And so on. This will be explored not only in the remainder of this section, but also throughout much of this book.

So, let us go into more detail now and develop the preceding abstract principle. Every set is a category, with the elements as the objects and the only arrows being the identity arrows of the elements. Thus, a set is a category with a most uninteresting structure, that is, there are no structural relations between different objects. In fact, the empty set $\emptyset$ also constitutes a category. This category has no objects and no arrows. This may strike you as the utmost triviality, but it turns out that for some formal constructions, it is quite useful to include this particular category.

However, reversely, we also have the category of sets, denoted by **Sets**, and also the category of finite sets. The objects of these categories are now sets, one of them being again the empty set $\emptyset$, and the morphisms are mappings

$$f : S_1 \to S_2 \tag{2.3.8}$$

between sets. In view of our above discussion of distinctions, this leads us to the concept of isomorphism:

Definition 2.3.2 Two objects A_1, A_2 of a category are *isomorphic* if there exist morphisms $f_{12} : A_1 \to A_2$, $f_{21} : A_2 \to A_1$ with

$$f_{21} \circ f_{12} = 1_{A_1}, \quad f_{12} \circ f_{21} = 1_{A_2}. \tag{2.3.9}$$

In this case, the morphisms f_{12}, f_{21} are called isomorphisms.

An *automorphism* of an object A is an isomorphism $f : A \to A$.

Of course, 1_A is an automorphism of A, but there may also exist others. Often, an automorphism is considered as a symmetry of A.

Since an automorphism can be inverted, the automorphisms of an object A of a category form a group, the automorphism group of A. In fact, this is how the group concept historically emerged. But we may then turn things around and consider a group as an abstract object that might be *represented* as the group of automorphisms of some object in a category. We'll return to that issue.

(2.3.9) means that isomorphisms are invertible morphisms. Isomorphic objects are then characterized by having the same morphisms, as follows from the associativity law. That is, for example when $f_{12} : A_1 \to A_2$ is an isomorphism, then a morphism $g : A_2 \to B$ corresponds to the morphism $g \circ f_{12} : A_1 \to B$, and similarly in other directions. In particular, 1_{A_2} then corresponds to f_{12}.

There may, however, exist more than one isomorphism between isomorphic objects A_1, A_2. In that case, the identification of the morphisms of

these two objects is not canonical as it depends on the choice of such an isomorphism. In fact, we can precompose an isomorphism $f_{12} : A_1 \to A_2$ with any automorphism $f_1 : A_1 \to A_1$ or postcompose it with any automorphism $f_2 : A_2 \to A_2$ to obtain another isomorphism. Conversely, whenever $f_{12}, g_{12} : A_1 \to A_2$ are two isomorphisms, then $g_{12}^{-1} \circ f_{12}$ is an automorphism of A_1, and $g_{12} \circ f_{12}^{-1}$ is an automorphism of A_2. Thus, the identification of two isomorphic objects is only determined up to the automorphisms of either of them. In fact, the automorphism groups of two isomorphic objects are themselves isomorphic. This is an instantiation of the fact that isomorphic objects have the same relational structure with other objects, in this case with themselves. Again, however, the automorphism groups may in turn possess certain symmetries, making this identification noncanonical again.

As observed, since automorphisms can be inverted, the automorphisms of an object A of a category form a group. In that sense, the concept of a morphism is a generalization of that of an automorphism in two ways. Firstly, it need not be invertible, and secondly, it need not map an object A to itself, but can map it to another object B of the same category. Morphisms can be composed. In distinction to a monoid or group where any two elements can be freely composed, however, here we have the restriction that the domain of the second morphism has to contain the codomain of the first one in order that they can be composed. Since in a monoid or group, all elements have the monoid or group itself as their domain and codomain, there is no such restriction for the composition of monoid or group elements. Thus, we had observed in Lemma 2.3.1 that the categories with a single object are precisely the monoids.

In any case, the most fundamental of the monoid or group laws, associativity, has to be preserved for the composition of morphisms. In a certain sense, associativity is a higher law, as it is about the composition of compositions. It stipulates that such a composition of compositions does not depend on the order in which we compose the compositions. This has to be distinguished from the property of commutativity which requires that a composition of group elements be independent of the order of these elements. Commutativity does not hold for a general monoid or group. Commutative monoids or groups constitute a special subclass of all monoids or groups, with many additional properties that are not shared by other monoids or groups in general.

Thus, a category can be considered as a directed graph, as some kind of generalized monoid, or as a set with some additional structure of directed relations between its elements.

Again, within a category, we cannot distinguish between isomorphic objects. We may thus wish to identify them, but need to keep in mind that such an identification need not be canonical as it depends on the choice of an isomorphism, as explained above. The important point here is that the objects of a category are determined only up to isomorphism. The view of category is that an object B of a category $\mathbf{C}$ is characterized by its relations with other objects, that is, by the sets $\mathrm{Hom}_{\mathbf{C}}(., B)$ and $\mathrm{Hom}_{\mathbf{C}}(B, .)$ of morphisms $f : A \to B$ and $g : B \to C$, respectively, and as we have seen, for isomorphic objects B_1 and B_2, the corresponding sets can

be identified, although not necessarily canonically. Thus, in a category, isomorphic objects cannot be distinguished by their relations with other objects.

In this sense, the category of finite sets contains just one single object for each $n \in \mathbb{N} \cup \{0\}$, the set with n elements, because any two sets with the same number of elements are isomorphic within the category of sets. Thus, the structure of the category of sets consists essentially in the cardinality. Again, this is notwithstanding the fact that the isomorphisms between sets of the same cardinality are not canonical as they can be composed with arbitrary permutations of the elements of the sets. In particular, the automorphism group of a set of n elements is the group $\mathfrak{S}_n$ of permutations of its elements introduced at the end of Sect. 2.1.6.

A poset becomes a category when we stipulate that there is an arrow $a \to b$ whenever $a \leq b$. In turn, we also have the category of posets, with arrows $m : A \to B$ between posets now given by monotone functions, that is, whenever $a_1 \leq a_2$ in A, then $m(a_1) \leq m(a_2)$ in B. Again, while we can consider a category as a graph, we can also consider the category of graphs. Morphisms are then mappings g between graphs $\Gamma_1 \to \Gamma_2$ that preserve the graph structure, that is, map edges to edges.

There can be categories with the same objects, but different morphisms. For instance, we can consider the category whose objects are sets, but whose morphisms are *injective* maps between sets. As another example, for a category with metric spaces as its objects, we could take the isometries as morphisms, that is, the mappings $f : (S_1, d_1) \to (S_2, d_2)$ with $d_2(f(x), f(y)) = d_1(x, y)$ for all $x, y \in S_1$. Alternatively, we can also take the more general class of distance nonincreasing maps, that is, those $g : (S_1, d_1) \to (S_2, d_2)$ with $d_2(g(x), g(y)) \leq d_1(x, y)$ for all $x, y \in S_1$. The isomorphisms of the category, however, are the same in either case. Algebraic structures also naturally fall into the framework of categories. Again, a single structure can be considered as a category, but we can also form the category of all structures of a given type. Thus, as already explained, for instance a monoid M or a group G yields the category with M or G as its only object and the multiplications by monoid or group elements as the morphisms. Thus, the monoid or group elements are not objects, but (endo)morphisms of this category.[7] In fact, for a group considered as a category, every morphism is then an isomorphism, because the group elements are invertible.

Considering monoid or group elements as morphisms, of course, reflects the general idea of a monoid or group as consisting of operations. We have already noted that the associativity law for monoids and groups is included in the definition of a category. In particular, the axioms for a category can also be considered as generalizations of the group axioms, as we do not require invertibility of the operations. Thus, the concept of a monoid is natural within category theory even though in general the concept of a group is more important than that of a monoid. In fact, a category with a single

[7]Alternatively, we could also consider the elements of a group or monoid as the objects of the corresponding category. The morphisms would again be the multiplications by elements. Thus, the classes of objects and morphisms would coincide.

object M is nothing but a monoid, where the composition of morphisms then defines the monoid multiplication. Thus, there are many morphisms from this single object to itself. Conversely, we have the categories of monoids, groups, finite groups, abelian groups, free (abelian) groups, Lie groups, etc. In such a category of groups, for instance, an object is again a group, but a morphism now has to preserve the group structure, that is, be a group homomorphism. We should be quite careful here. A monoid M or group G considered as a category is not a subcategory[8] of the category **Monoids** of monoids or **Groups** of groups, resp. The reason is that in those two cases, the notion of a morphism is different. For a single group as a category, the multiplication by any group element as an operation on the group itself is a morphism. Within the category of groups, however, a morphism $\chi : G_1 \to G_2$ between two objects has to preserve their group structure. In particular, χ has to map the neutral element of G_1 to the neutral element of G_2. Analogously, of course, for the case of monoids.

There is a generalization of the foregoing. Let M again be a fixed monoid, with its unit element denoted by e, and with the product of the elements m, n simply written as mn. By definition, the category $\mathbf{B}M = M - \mathbf{Sets}$ consists of all representations of M, that is, of all sets X with an operation of M on X, i.e.,

$$\begin{aligned} &\mu : M \times X \to X \\ &\quad (m, x) \quad \mapsto mx \end{aligned}$$

$$\text{with } ex = x \text{ and } (mn)x = m(nx) \text{ for all } x \in X, m, n \in M. \qquad (2.3.10)$$

A morphism $f : (X, \mu) \to (Y, \lambda)$ then is a map $f : X \to Y$ which is equivariant w.r.t. the representations, that is,

$$f(mx) = mf(x) \text{ for all } m \in M, x \in X \qquad (2.3.11)$$

(where we also write $\lambda(m, y) = my$ for the representation λ). Expressed more abstractly,

$$f(\mu(m, x)) = \lambda(m, f(x)). \qquad (2.3.12)$$

For instance, when L is a left ideal of M, then left multiplication by M on L yields such a representation.

Another interpretation of a category, which leads us into logic, a topic to be taken up in Sect. 9.3, is that of a deductive system. The objects of a deductive system are interpreted as formulas, the arrows as proofs or deductions, and the operations on arrows as rules of inference. For formulas X, Y, Z and deductions $f : X \to Y, g : Y \to Z$, we have the binary operation of composition, yielding $g \circ f : X \to Z$, as an inference rule. Thus, by stipulating an equivalence relation for proofs, a deductive system becomes a category. Or putting it the other way around, a category is a formal encoding of a deductive system. See [74] and Sect. 9.3.

We now develop some general concepts.

Definition 2.3.3 An arrow $f : A \to B$ between two objects of a category **C** is called

[8] An obvious definition: A category **D** is a subcategory of the category **C** if every object D and every morphism $D_1 \to D_2$ of **D** is also an object or a morphism, resp., of **C**.

- a *monomorphism*, or shortly, *monic*, in symbols,

$$f : A \rightarrowtail B, \text{ or } f : A \hookrightarrow B, \tag{2.3.13}$$

  if for any morphisms $g_1, g_2 : C \to A$ in $\mathbf{C}$, $fg_1 = fg_2$ implies $g_1 = g_2$,
- an *epimorphism*, or shortly, *epic*, in symbols,

$$f : A \twoheadrightarrow B, \tag{2.3.14}$$

  if for any morphisms $h_1, h_2 : B \to D$ in $\mathbf{C}$, $h_1 f = h_2 f$ implies $h_1 = h_2$.

These notions generalize those of injective and surjective mappings between sets, as introduced in Sect. 2.1.2.

An isomorphism is both monic and epic. In the category **Sets** the converse also holds, that is a monic and epic morphism is an isomorphism (in short jargon: In **Sets**, monic epics are iso). In a general category, this need not be true, however. For instance, in the category of free abelian groups, $f : \mathbb{Z} \to \mathbb{Z}, n \mapsto 2n$ is monic and epic, but not iso.

The above definition is an instance of a general principle in category theory, to define properties through relations with other objects or morphisms within a category. We shall systematically explore this principle below.

Definition 2.3.4 A morphism f is called an *endomorphism* if its codomain coincides with its domain A, in symbols

$$f : A \circlearrowleft . \tag{2.3.15}$$

Thus, an automorphism is an invertible endomorphism.

Definition 2.3.5 A *subobject* A of the object B of the category $\mathbf{C}$ is a monomorphism $f : A \rightarrowtail B$.

Often, the monomorphism f is clear from the context, and we then simply call A a subobject of B.

Thus, for instance, in the category **Sets**, we can speak of subsets, whereas in the category **Groups**, we have subgroups. For the set $\{1, 2, \ldots, n\}$ of n elements, any collection $\{i_1, i_2, \ldots, i_m\}$ for any distinct $i_k \in \{1, 2, \ldots, n\}$, $m < n$, yields a subset. And the group $\mathfrak{S}_m$, introduced at the end of Sect. 2.1.6, of permutations of those m elements is a subgroup of $\mathfrak{S}_n$. .

$\mathfrak{S}_n$

The observation at the end of Sect. 2.1.6 that for a finite group G, the left translation l_g by any element $g \in G$ yields a permutation of the elements of G, with different elements inducing different permutations, means that we can consider G as a subgroup of the group of permutations of its elements. Slightly more abstractly: Every finite group is a subgroup of a symmetric group. This is known as Cayley's Theorem.

We can also consider the morphisms of one category $\mathbf{C}$ as the objects of another category $\mathbf{D}$. In other words, operations within one category can become the objects in another one. In particular, what we mean by an "object" in mathematics has little, if anything, to do with what an "object" is in ordinary language. In category theory, an object is anything on which we can perform systematic operations that relate it to other objects. And

when and since we can also operate on operations, they in turn can become objects.

But if we take operations as objects, what then are the operations of that category? The morphisms of **D** are arrows between morphisms of **C**, that is,

$$F : (f : A \to B) \to (g : C \to D), \tag{2.3.16}$$

given by a pair

$$\phi : A \to C, \psi : B \to D \tag{2.3.17}$$

of morphisms of **C** with

$$\psi \circ f = g \circ \phi. \tag{2.3.18}$$

One also expresses this relation by saying that the diagram

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\psi} \\ C & \xrightarrow{g} & D \end{array} \tag{2.3.19}$$

commutes. In short, the morphisms of a category of morphisms are commuting diagrams. As an example, when $f : A \to B$ is a morphism, the identity 1_f is obtained from the identities 1_A and 1_B through such a commutative diagram

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle 1_A}\downarrow & & \downarrow{\scriptstyle 1_B} \\ A & \xrightarrow{f} & B\,. \end{array} \tag{2.3.20}$$

We shall now derive a simple condition for constructing a commutative diagram from two morphisms. This is the basic situation considered in [94]. Let X, Y be objects in some category **C**, and let $\pi : X \to Y$ be a morphism. When $\mathbf{C} = \mathbf{Sets}$, that is, X, Y are sets and π is a map between them, then π defines an equivalence relation,

$$x_1 \sim x_2 \text{ iff } \pi(x_1) = \pi(x_2). \tag{2.3.21}$$

Let $f : X \to X$ be another morphism. We can then find a morphism $\tilde{f} : Y \to Y$ for which the diagram

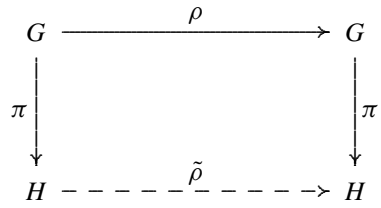
$$\begin{array}{ccc} X & \xrightarrow{\quad f \quad} & X \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ Y & \overset{\tilde{f}}{\dashrightarrow} & Y \end{array} \tag{2.3.22}$$

commutes iff f commutes with the equivalence relation (2.3.21), that is,

$$\text{if } x_1 \sim x_2 \text{ then also } f(x_1) \sim f(x_2). \tag{2.3.23}$$

When **C** is a category of sets with additional structure, then condition (2.3.23) becomes simpler to check, because we can utilize the additional structure. For instance, when $\mathbf{C} = \mathbf{Groups}$ and $\pi : G \to H$ is a group

homomorphism, then for a homomorphism $\rho : G \to G$, we get a commutative diagram



(2.3.24)

iff

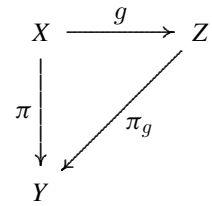
$$\rho(\ker \pi) \subset \ker \pi, \tag{2.3.25}$$

because in that case, (2.3.25) implies (2.3.23).

The condition (2.3.23) is not a categorical one. Therefore, we now reformulate it in a more abstract manner. Given the morphism $\pi : X \to Y$ in our category $\mathbf{C}$, consider the set of morphisms

$$K(\pi) = \{g : X \to Z \text{ morphism of } \mathbf{C} : \pi \text{ factors through } g\}, \tag{2.3.26}$$

that is, where there exists a morphism $\pi_g : Z \to X$ with $\pi = \pi_g \circ G$. Thus, $g \in K(\pi)$ iff there exists a commutative diagram

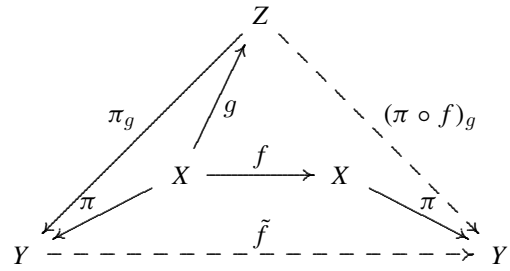


(2.3.27)

Lemma 2.3.2 *A necessary and sufficient condition for (2.3.22) to commute is*

$$K(\pi) \subset K(\pi \circ f). \tag{2.3.28}$$

Proof We consider the following diagram

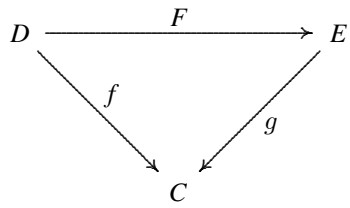


(2.3.29)

Let $g \in K(\pi)$. When (2.3.22) commutes, we have $\tilde{f}$ and can put $(\pi \circ f)_g = \tilde{f} \circ \pi_g$. Thus, $g \in K(\pi \circ f)$. For the other direction, we simply observe that $\pi \in K(\pi)$ with $Z = Y$ and $\pi_\pi = \mathrm{id}_Y$. Then, $\tilde{f} = (\pi \circ f)_\pi$ lets (2.3.22) commute. □

In the case of groups, as in (2.3.24), (2.3.25), the crucial g is the projection onto the quotient group $G/\ker\pi$.

We can also fix an object C of $\mathbf{C}$ and consider the category $\mathbf{C}/C$ of all morphisms $f : D \to C$ from objects D of $\mathbf{C}$. This category is called a slice or comma category. A morphism $f \to g$ between two objects of this slice category, that is, from an arrow $f : D \to C$ to an arrow $g : E \to C$ is then a commutative diagram



(2.3.30)

that is, an arrow $F : D \to E$ with $f = g \circ F$.

We can then also go ahead and form categories $\mathcal{C}$ of categories.[9] That is, the objects of $\mathcal{C}$ are categories $\mathbf{C}$, and the morphisms $F : \mathbf{C} \to \mathbf{D}$ of $\mathcal{C}$, called *functors*, then preserve the category structure. This means that they map objects and arrows of $\mathbf{C}$ to objects and arrows of $\mathbf{D}$, satisfying

$$F(f : A \to B) \text{ is given by } F(f) : F(A) \to F(B) \tag{2.3.31}$$

$$F(g \circ f) = F(g) \circ F(f) \tag{2.3.32}$$

$$F(1_A) = 1_{F(A)} \tag{2.3.33}$$

for all A, B, f, g. Thus, the image of an arrow under F is an arrow between the images of the corresponding objects (domain and codomain) under F, preserving compositions, and mapping identities to identities.

Functors play a very important role as typically one wants to assign to objects of a category with a perhaps complicated structure objects of a category with less structure that nevertheless capture the important qualitative features of the former. For instance, we can associate to a topological space its cohomology groups, as will be explained below. These groups are algebraic objects that encode qualitative topological properties of these spaces. The question that typically arises from such constructions is whether they capture all the relevant features. In the present example, this leads to the question to what extent the cohomology groups determine the topology of a space.

In general, a functor that maps one category to another one with less structure is called forgetful.

[9] In fact, these form more naturally so-called bicategories. We suppress this technical point here, however. See for instance [80]. More importantly, one has to be careful to avoid paradoxes of self-reference. Therefore, one respects the axioms of set theory as listed in Sect. 2.2 and considers only sets from a universe U. A category will be called small if both its objects and its arrows constitute a set from U (see Definition 8.1.2). One then looks only at categories of small categories.

Given two categories $\mathbf{C}, \mathbf{D}$, we can then also look at the category $\mathbf{Fun}(\mathbf{C}, \mathbf{D})$ of all functors $F : \mathbf{C} \to \mathbf{D}$. The morphisms of this category are called *natural transformations*. Thus, a natural transformation

$$\theta : F \to G \tag{2.3.34}$$

maps a functor F to another functor G, preserving the structure of the category $\mathbf{Fun}(\mathbf{C}, \mathbf{D})$. What is that structure, and how can it be preserved? Well, the defining property of a functor is that it maps morphisms of $\mathbf{C}$ to morphisms of $\mathbf{D}$. Thus, given a morphism $f : C \to C'$ in $\mathbf{C}$, we obtain morphisms $Ff : FC \to FC'$ and $Gf : GC \to GC'$ in $\mathbf{D}$. A natural transformation $\theta : F \to G$ then has to respect that relation. That means that for each $C \in \mathbf{C}$, it induces a morphism

$$\theta_C : FC \to GC \tag{2.3.35}$$

such that the diagram

$$\begin{array}{ccc} FC & \xrightarrow{\theta_C} & GC \\ {\scriptstyle Ff}\downarrow & & \downarrow{\scriptstyle Gf} \\ FC' & \xrightarrow{\theta_{C'}} & GC' \end{array} \tag{2.3.36}$$

commutes.

As will be investigated in more detail in Sect. 8.3, in particular, we can consider functor categories of the form $\mathbf{Sets}^{\mathbf{C}}$, involving the category $\mathbf{Sets}$ of sets and some small category $\mathbf{C}$ ($\mathbf{C}$ is called small if its collections of objects and of arrows are both sets, see Definition 8.1.2). The objects of $\mathbf{Sets}^{\mathbf{C}}$ are functors

$$F, G : \mathbf{C} \to \mathbf{Sets}, \tag{2.3.37}$$

and its arrows are natural transformations

$$\phi, \psi : F \to G. \tag{2.3.38}$$

According to (2.3.35) and (2.3.36), this means that, for instance, $\phi : F \to G$ for each $C \in \mathbf{C}$ has to induce a morphism

$$\phi_C : FC \to GC \tag{2.3.39}$$

such that the diagram

$$\begin{array}{ccc} FC & \xrightarrow{\phi_C} & GC \\ {\scriptstyle Ff}\downarrow & & \downarrow{\scriptstyle Gf} \\ FC' & \xrightarrow{\phi_{C'}} & GC' \end{array} \tag{2.3.40}$$

commutes.

We also need the opposite category $\mathbf{C}^{\mathbf{op}}$, obtained from $\mathbf{C}$ by taking the same objects, but reversing the direction of all arrows. This simply means that each arrow $C \to D$ in $\mathbf{C}^{\mathbf{op}}$ corresponds to an arrow $D \to C$ in $\mathbf{C}$. In some cases, this procedure is quite natural. For instance, when the category $\mathbf{C}$ is a poset, this simply amounts to replacing $x \leq y$ by $y \geq x$.

We can now consider the category $\mathbf{Sets}^{\mathbf{C}^{\mathbf{op}}}$ for some fixed small category $\mathbf{C}$. Here, for instance, the category $\mathbf{C}$ could be $\mathcal{P}(S)$, the objects of which

are the subsets U of some set S and the morphisms the inclusions $V \subset U$. That is, $\mathcal{P}(S)$ has the structure of a poset, with the ordering relation $\leq$ given by inclusion $\subset$. This poset has a largest element, S, and a smallest one, $\emptyset$. (In fact, $\mathcal{P}(S)$ is a Boolean algebra, with the operations of intersection and union.) This category will be treated in Sect. 2.4.

From the preceding, we may form the impression that the most important example of a category is the category **Sets** of sets. Much of the terminology is oriented towards that example, for instance the representation of morphisms by arrows reminds us of maps between sets. The objects of many other categories are sets with some additional structure, such as a partial order or a group structure. We then have a natural functor from such a category to **Sets**, the so-called forgetful functor that simply forgets that additional structure. Also, it is a fundamental principle to be explored below in more detail that an object of a category can to a large extent be described or characterized by its hom-set, that is, by the set of morphisms into that object or the morphisms from that object.

Another basic example is the category of groups, **Groups**, which we have already discussed above. We have pointed out that the notion of a morphism is inspired by that of a homomorphism between groups. Homomorphisms between groups preserve the group structure, and a morphism between objects of some category has to preserve the characteristic structure of that category.

In order to show the applicability or limitations of the concepts introduced below, we shall often consider other categories. For that purpose, a particularly useful example will be the one consisting of a single poset. Here, the objects are the elements of that poset. As already emphasized, this should not be confused with the category of all posets, where the objects are posets themselves, instead of elements of posets.

The topic of categories will be taken up systematically below in Chap. 8. In order to appreciate the general approach of that chapter, it will be useful to first look at some more particular mathematical structures in depth. We shall do this in the following chapters.

Before moving on in this direction, let us insert a small warning. Even though category theory provides us with abstract principles and constructions that apply simultaneously to all categories, nevertheless, the concrete content of such constructions might be rather different according to the type of category under consideration. On one hand, we have the categories whose objects are simply elements of some set. These elements do not possess any internal structure. They may stand in binary relations F, and such relations then define the morphisms. In a set, there are no nontrivial such relations, that is, $F(s_1, s_2) = 1$ only for $s_1 = s_2$. In a poset, such a relation is denoted by $\leq$, that is, $s_s \leq s_2$ iff $F(s_1, s_2) = 1$. Such binary relations can also be represented geometrically as digraphs, that is, we draw an edge from s_1 to s_2 whenever $F(s_1, s_2) = 1$.

In contrast, at the next level, we have categories, like **Sets**, **Posets** or **Groups** whose objects do have some particular internal structure (although trivial in the case of **Sets**). Morphisms are required to preserve that structure, that is, be structure homomorphisms. For this type of categories, the constructions of category theory will turn out to much more useful than

for the preceding ones, whose objects are elements without internal structure. While the preceding categories may serve to illustrate some of those constructions, as examples of the general thrust of the theory they may be somewhat misleading.

2.4 Presheaves

In this section, we shall take a first glimpse at the concept of a presheaf, which will be treated in more detail and depth in Sects. 4.5 and 8.4 below.

Here is a preliminary definition that will be refined in Sect. 4.5 where we shall work in the category of topological spaces instead of sets. A bundle over the set S is a surjective map $p : T \to S$ from some other set T. For $x \in S$, the preimage $p^{-1}(x)$ is called the fiber over x. S is called the base space, and T is the total space. A section of such a bundle is a map $s : S \to T$ with $p \circ s = 1_S$. Thus, a section associates to each element x of the base space an element of the fiber over x. Usually, for a given bundle, the space of sections is constrained; that is, not every such s with $p \circ s = 1_S$ represents a valid section of the bundle, but the sections need to satisfy certain restrictions or constraints.

Here is an interpretation. S could be a set of properties, observables or features of possible objects, like size, color, texture, material. The fiber over such a feature $x \in S$ then contains the possible values that this feature can assume. When x stands for color, the fiber over x might contain the values 'red', 'green', 'blue' etc., or if we desire, also more precise shades of color like 'azure blue', 'crimson', or 'yellowish pink'. A section then assigns to each object the values of its properties, that is, in the current example, its color, size, texture, material etc. The whole point now is that the values of the various properties in general are not independent of each other. A trivial example might elucidate this issue. When the object in question is a piece of gold, then the value 'gold' for the material constrains the color to be 'golden', and also its size and texture will obey certain restrictions. Deeper and more interesting examples come from theoretical biology, and lead, in fact, to the core of the field of morphology. Two hundred years ago, the biologist Cuvier (1769–1832), the founder of scientific paleontology and comparative anatomy, had already emphasized that a plant or animal does not just consist of an arbitrary collection of feature values, but that those are highly interdependent and determined by its mode of living. According to his principle of "Correlation of parts", the anatomical structures of the various organs of an animal are functionally related to each other and the structural and functional characteristics of the organs are all derived from the particular mode of living of the animal within its environment. Mammals are not only viviparous, but also typically possess fur and characteristic anatomical features, and carnivores not only have teeth and jaws adapted to catch and handle their pray, but also digestive tracts suited for their meat diets, and feet depending on the way they chase their prey, and so on. In fact, based on such correspondences, he could perform the stunning feat of reconstructing a particular dinosaur on the sole basis

of a very incomplete fossil consisting only of a claw. Later on, when a more complete fossil skeleton was found, it agreed in remarkable detail with Cuvier's reconstruction which made him quite famous.[10] Translated into our language of bundles and sections, there exist strong correlations, constraints and restrictions between the values of the various features, and knowing such constraints, one can reconstruct much of a section from the values of particular features. The biological aspects will be explored in detail elsewhere. Here, we want to use this to introduce a key concept, that of a presheaf.

We recall the functor category $\mathbf{Sets}^{\mathbf{C}^{\mathbf{op}}}$ for some small category $\mathbf{C}$ where $\mathbf{C}^{\mathbf{op}}$ is obtained from $\mathbf{C}$ by reversing the directions of all arrows, and stipulate

Definition 2.4.1 An element P of $\mathbf{Sets}^{\mathbf{C}^{\mathbf{op}}}$ is called a *presheaf* on $\mathbf{C}$.

For an arrow $f : V \to U$ in $\mathbf{C}$, and $x \in PU$, the value $Pf(x)$, where $Pf : PU \to PV$ is the image of f under P, is called the *restriction* of x along f.

Thus, a presheaf formalizes the possibility of restricting collections of objects, that is, the—possibly structured—sets assigned to subsets of S, from a set to its subsets.

We can put this into the more general context that will be developed in Chap. 8, but this will not be indispensable for understanding the current section. Anticipating some of Sect. 8.4 and also of Sect. 8.3, we consider $\mathrm{Hom}_{\mathbf{C}}(V, U)$, the set of morphisms in the category $\mathbf{C}$ from the object V to the object U. Each object $U \in \mathbf{C}$ then yields the presheaf yU defined on an object V by

$$yU(V) = \mathrm{Hom}_{\mathbf{C}}(V, U) \tag{2.4.1}$$

and on a morphism $f : W \to V$ by

$$\begin{aligned} yU(f) : \quad \mathrm{Hom}_{\mathbf{C}}(V, U) &\to \mathrm{Hom}_{\mathbf{C}}(W, U) \\ h &\mapsto h \circ f. \end{aligned} \tag{2.4.2}$$

We shall also see in Sect. 8.3 that when $f : U_1 \to U_2$ is a morphism of $\mathbf{C}$, we obtain a natural transformation $yU_1 \to yU_2$ by composition with f, so that we get the Yoneda embedding (Theorem 8.3.1)

$$y : \mathbf{C} \to \mathbf{Sets}^{\mathbf{C}^{\mathbf{op}}}. \tag{2.4.3}$$

The presheaf yU, with

$$yU(V) = \mathrm{Hom}_{\mathbf{C}}(V, U), \tag{2.4.4}$$

is also called the functor of points as it probes U by morphisms from other members V of the category. When we work with the category **Sets** and V is a single element set, then any morphism from such a single element set to a set U determines an element in U, a point of U. When V is a more general set, then this yields, in naive terminology, a family of points in U parametrized by V. The categorial approach thus naturally incorporates such generalized

[10] We refer to [44] for a conceptual analysis of Cuvier's position in the history of biology.

points. For instance, when we are in the category of algebraic varieties (to be defined below), we probe an algebraic variety U by considering morphisms from other algebraic varieties V, be they classical points or more general varieties.

When we go in the opposite direction and consider

$$zU(V) = \mathrm{Hom}_{\mathbf{C}}(U, V), \tag{2.4.5}$$

we obtain what is called the functor of functions. Here, classically, one would take as V a field such as the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$.

Of course, we can then also let U and V vary simultaneously, to make the construction symmetric.

We now return to the category $\mathcal{C} = \mathcal{P}(S)$ of subsets of some set S. For a presheaf $P : \mathcal{P}(S)^{\mathrm{op}} \to \mathbf{Sets}$, we then have the restriction maps

$$p_{VU} : PV \to PU \text{ for } U \subset V \tag{2.4.6}$$

that satisfy

$$p_{UU} = 1_{PU} \tag{2.4.7}$$

and

$$p_{WU} = p_{VU} \circ p_{WV} \text{ whenever } U \subset V \subset W. \tag{2.4.8}$$

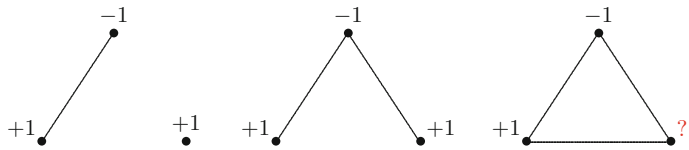
Definition 2.4.2 The presheaf $P : \mathcal{P}(S)^{\mathrm{op}} \to \mathbf{Sets}$ is called a *sheaf* if it satisfies the following condition. If $U = \bigcup_{i\in I} U_i$ for some family $(U_i)_{i\in I} \subset \mathcal{P}(S)$ and $\pi_i \in PU_i$ satisfies $p_{U_i,U_i\cap U_j}\pi_i = p_{U_j,U_i\cap U_j}\pi_j$ for all $i, j \in I$, then there exists a unique $\pi \in PU$ with $p_{UU_i}\pi = \pi_i$ for all i.

Thus, whenever the π_i are compatible in the sense that the restrictions of π_i and π_j to $U_i \cap U_j$ always agree, then they can be patched together to an element π of PU that restricts to π_i on PU_i.

The interpretation of a presheaf over $\mathcal{P}(S)$ that is most important for our current purposes takes place within the framework of fiber bundles developed at the beginning of this section. The set assigned to some $U \subset S$ by the presheaf P then would simply be the set of sections over U of a fiber bundle with base S. Such sections over a subset U are called local because they need not extend to all of S. Sections defined over all of S could be called global, and thus, local sections need not admit extensions to global sections. In contrast, by the presheaf condition, whenever we have a local section over some U, we can restrict it to any $V \subset U$. The sheaf condition, in turn, stipulates that locally compatible local sections can be patched together to a global section. Not every presheaf is a sheaf, and so, such an extension from local compatibility to what one may call global coherence need not always be possible.

For instance, when our fiber bundle is a Cartesian product $S \times \mathbb{R}$ or $S \times \mathbb{C}$, then all fibers are $\mathbb{R}$, and a local section over U is nothing but a real valued function on U. A presheaf then might stipulate further conditions or restrictions on such functions. In particular, when S carries some additional structure, like that of a metric, we could require the functions to respect

that structure. In the case where S is equipped with a metric $D(.,.)$, we might request, for instance, that for every local section over U, that is, for every function $f : U \to \mathbb{R}$ that belongs to the presheaf, we have $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in U$. Or when S is a graph, we could only admit functions with values ± 1 and require that any such function on a subgraph U assign different values to neighboring vertices. That is, when $f(x) = 1$ for some vertex x, then $f(y) = -1$ for all neighbors y of x, and conversely. Such a function then cannot be extended to any subgraph that contains a triangle, that is, three vertices x, y, z each of which is a neighbor of the other two. More generally and precisely, a graph is called bipartite iff it admits such a function f. Bipartite graphs do not contain triangles, nor other cycles of odd length.[11] They consist of two classes of vertices, so that a function f with the above properties assigns the value +1 to one class and -1 to the other.



Here, we have sections for the first two graphs, but none for the third. This phenomenon is also called frustration. Of course, the first two graphs are subgraphs of the third, and so, we see here an example where a local section cannot be extended to a global section.

Returning to the biological principle of the "Correlation of parts", the sections of a presheaf would correspond to the different species, and the fact that only specific value combinations in the various fibers are realized by some section then reflects those correlations and constraints.

We now describe and further explore a different biological realization of presheafs, and discuss a proposal of Benecke and Lesne [11] to describe genomes in this formal framework. The DNA of a cell of a member of a biological species is a linear string composed of instances of four nucleotides, labelled by the letters A, T, C, G.[12] That is, we have a finite sequence (with about 3 billion elements in the case of humans) of positions, called genetic loci, each of which is filled by one of the letters A, T, C, G. Looking at this from the point of view of topology, we take as base space the space of genetic loci of a species, with a metric given by the distance between positions in the linear arrangement.[13] As the fiber, we take the possible nucleotide values, that is, A, T, C, and G. This fiber then carries a natural probability measure (see Sect. 4.4 below for the formal definition) given by the relative frequencies of the nucleotide values. In fact, first the fiber over each locus carries such a measure. We can then also compare fibers

[11] see Sect. 3.4.1 for the definition of a cycle.

[12] With certain exceptions that need not concern us here, all the cells of a given organism carry the same DNA sequence.

[13] We assume here that there is a one-to-one correspondence between the loci of different members of the species. That is, we assume that the only differences between individuals are given by point mutations, but not by insertions or deletions of nucleotide strings in the genome.

over different loci by the distance between those measures.[14] We can also simply look at the abstract fiber of the four nucleotide values and obtain a measure on it by averaging over all genetic loci.

An individual genome is then a section of this fiber bundle. The space of sections then yields a sheaf, even though, of course, not every section needs to be realized by the genome of some individual. Again, spaces of genomes then induce measures on the space of sections as well as on the individual fibers, or more generally, on the collection of fibers over any given set of loci. When the set of loci of two populations is in 1-1 correspondance, we can then look at the distance between the measures induced on the space of sections by the two populations. We can then define the genetic distance between the populations as such a distance between measures on the space of sections.

2.5 Dynamical Systems

Definition 2.5.1 A *dynamical system* is a homomorphism ϕ from the additive group $\mathbb{Z}$ of integers or real numbers $\mathbb{R}$ (or the semigroup of nonnegative integers or reals) to the group of (invertible) selfmaps $\mathcal{F}(S)$ of some set S. When the domain is the group (monoid) of (nonnegative) integers, we speak of a discrete-time dynamical system, and when we have the (nonnegative) reals, we talk about a continuous-time system.

In particular, $0 \in \mathbb{R}$ is mapped to id_S, and $\phi(t_1 + t_2) = \phi(t_1) \circ \phi(t_2)$. Often, there is more structure; for instance, S could be a topological space (see Definition 4.1.1) or a differentiable manifold (see Definition 5.3.3), and the selfmaps could be homeomorphisms (see Definition 4.1.11) or diffeomorphisms. Or, S could be a vector space, and the maps linear. We can write this as a diagram

$$\begin{array}{ccc} t_1 & \longrightarrow & t_2 \\ \downarrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \phi} \\ \phi(t_1) & \longrightarrow & \phi(t_2) \end{array} \tag{2.5.1}$$

The variable t here is considered as time. The value $\phi(0)$ is called the *initial value* of the dynamical system (2.5.1). One usually writes $\phi(x, t)$ for $\phi(t)(x)$, the value of the map $\phi(t)$ applied to the initial value $x = \phi(0)$. This expresses the dependence on time t and the initial value x. The collection of points $\phi(x, t)$ as t varies is called the *orbit* of x.

[14] Here, we can use the distance induced by the Fisher metric on the space of measures. We can also utilize the Kullback-Leibler divergence, which is not quite a distance, in fact, because it is not symmetric. For the definition and for a geometric view of these distances, see [3, 6].

In the discrete-time case, we can simply take a map $F : S \to S$ (invertible when the system is to be defined on $\mathbb{Z}$) and consider its iterates, that is, put

$$\phi(x, n) = F^n(x), \tag{2.5.2}$$

the n-fold iterate of F (when $n < 0$ and F is therefore assumed to be invertible, $F^n = (F^{-1})^{-n}$). Thus, $\mathbb{N}$ or $\mathbb{Z}$ operates on S by iteration of a self-map. As a diagram, this looks like

$$\begin{array}{ccccccc}
 & \longrightarrow & n-1 & \xrightarrow{+1} & n & \xrightarrow{+1} & n+1 \longrightarrow \\
\cdots & & \downarrow & & \downarrow & & \downarrow \quad \cdots \\
 & \longrightarrow & F^{n-1}(x) & \xrightarrow{F} & F^n(x) & \xrightarrow{F} & F^{n+1}(x) \longrightarrow
\end{array} \tag{2.5.3}$$

Such a discrete-time dynamical system, that is, a set S equipped with an endomorphism F, is also called an automaton.

http://www.springer.com/978-3-319-20435-2

Mathematical Concepts
Jost, J.
2015, XV, 312 p. 130 illus., 16 illus. in color., Softcover
ISBN: 978-3-319-20435-2