CHAPTER 1

Categories, Functors, Natural Transformations

> Frequently in modern mathematics there occur phenomena of "naturality".
>
> Samuel Eilenberg and Saunders Mac Lane, "Natural isomorphisms in group theory" [**EM42b**]

A **group extension** of an abelian group H by an abelian group G consists of a group E together with an inclusion of $G \hookrightarrow E$ as a normal subgroup and a surjective homomorphism $E \twoheadrightarrow H$ that displays H as the quotient group E/G. This data is typically displayed in a diagram of group homomorphisms:

$$0 \to G \to E \to H \to 0.$$[1]

A pair of group extensions E and E' of G and H are considered to be equivalent whenever there is an isomorphism $E \cong E'$ that *commutes with* the inclusions of G and quotient maps to H, in a sense that is made precise in §1.6. The set of equivalence classes of *abelian* group extensions E of H by G defines an abelian group $\mathrm{Ext}(H, G)$.

In 1941, Saunders Mac Lane gave a lecture at the University of Michigan in which he computed for a prime p that $\mathrm{Ext}(\mathbb{Z}[\frac{1}{p}]/\mathbb{Z}, \mathbb{Z}) \cong \mathbb{Z}_p$, the group of p-adic integers, where $\mathbb{Z}[\frac{1}{p}]/\mathbb{Z}$ is the Prüfer p-group. When he explained this result to Samuel Eilenberg, who had missed the lecture, Eilenberg recognized the calculation as the homology of the 3-sphere complement of the p-adic solenoid, a space formed as the infinite intersection of a sequence of solid tori, each wound around p times inside the preceding torus. In teasing apart this connection, the pair of them discovered what is now known as the **universal coefficient theorem** in algebraic topology, which relates the *homology* H_* and *cohomology groups* H^* associated to a space X via a group extension [**ML05**]:

$$0 \to \mathrm{Ext}(H_{n-1}(X), G) \to H^n(X, G) \to \mathrm{Hom}(H_n(X), G) \to 0\,. \tag{1.0.1}$$

To obtain a more general form of the universal coefficient theorem, Eilenberg and Mac Lane needed to show that certain isomorphisms of abelian groups expressed by this group extension extend to spaces constructed via direct or inverse limits. And indeed this is the case, precisely because the homomorphisms in the diagram (1.0.1) are *natural* with respect to continuous maps between topological spaces.

The adjective "natural" had been used colloquially by mathematicians to mean "defined without arbitrary choices." For instance, to define an isomorphism between a finite-dimensional vector space V and its **dual**, the vector space of linear maps from V to the

[1] The zeros appearing on the ends provide no additional data. Instead, the first zero implicitly asserts that the map $G \to E$ is an inclusion and the second that the map $E \to H$ is a surjection. More precisely, the displayed sequence of group homomorphisms is **exact**, meaning that the kernel of each homomorphism equals the image of the preceding homomorphism.

ground field $\Bbbk$, requires a choice of basis. However, there is an isomorphism between V and its double dual that requires no choice of basis; the latter, but not the former, is *natural*.

To give a rigorous proof that their particular family of group isomorphisms extended to inverse and direct limits, Eilenberg and Mac Lane sought to give a mathematically precise definition of the informal concept of "naturality." To that end, they introduced the notion of a *natural transformation*, a parallel collection of homomorphisms between abelian groups in this instance. To characterize the source and target of a natural transformation, they introduced the notion of a *functor*.[2] And to define the source and target of a functor in the greatest generality, they introduced the concept of a *category*. This work, described in "The general theory of natural equivalences" [**EM45**], published in 1945, marked the birth of category theory.

While categories and functors were first conceived as auxiliary notions, needed to give a precise meaning to the concept of naturality, they have grown into interesting and important concepts in their own right. Categories suggest a particular perspective to be used in the study of mathematical objects that pays greater attention to the maps between them. Functors, which translate mathematical objects of one type into objects of another, have a more immediate utility. For instance, the Brouwer fixed point theorem translates a seemingly intractable problem in topology to a trivial one ($0 \neq 1$) in algebra. It is to these topics that we now turn.

Categories are introduced in §1.1 in two guises: firstly as universes categorizing mathematical objects and secondly as mathematical objects in their own right. The first perspective is used, for instance, to define a general notion of *isomorphism* that can be specialized to mathematical objects of every conceivable variety. The second perspective leads to the observation that the axioms defining a category are self-dual.[3] Thus, as explored in §1.2, for any proof of a theorem about all categories from these axioms, there is a dual proof of the dual theorem obtained by a syntactic process that is interpreted as "turning around all the arrows."

Functors and natural transformations are introduced in §1.3 and §1.4 with examples intended to shed light on the linguistic and practical utility of these concepts. The category-theoretic notions of *isomorphism*, *monomorphism*, and *epimorphism* are invariant under certain classes of functors, including in particular the *equivalences of categories*, introduced in §1.5. At a high level, an equivalence of categories provides a precise expression of the intuition that mathematical objects of one type are "the same as" objects of another variety: an equivalence between the category of matrices and the category of finite-dimensional vector spaces equates high school and college linear algebra.

In addition to providing a new language to describe emerging mathematical phenomena, category theory also introduced a new proof technique: that of the diagram chase. The introduction to the influential book [**ES52**] presents *commutative diagrams* as one of the "new techniques of proof" appropriate for their axiomatic treatment of homology theory. The technique of diagram chasing is introduced in §1.6 and applied in §1.7 to construct new natural transformations as *horizontal* or *vertical composites* of given ones.

[2] A brief account of functors and natural isomorphisms in group theory appeared in a 1942 paper [**EM42b**].

[3] As is the case for the duality in projective plane geometry, this duality can be formulated precisely as a feature of the first-order theories that axiomatize these structures.

1.1. Abstract and concrete categories

> It frames a possible template for any mathematical theory: the theory should have *nouns* and *verbs*, i.e., objects, and morphisms, and there should be an explicit notion of composition related to the morphisms; the theory should, in brief, be packaged by a category.
>
> Barry Mazur, "When is one thing equal to some other thing?" [**Maz08**]

DEFINITION 1.1.1. A **category** consists of

- a collection of **objects** $X, Y, Z, \ldots$
- a collection of **morphisms** $f, g, h, \ldots$

so that:

- Each morphism has specified **domain** and **codomain** objects; the notation $f\colon X \to Y$ signifies that f is a morphism with domain X and codomain Y.
- Each object has a designated **identity morphism** $1_X\colon X \to X$.
- For any pair of morphisms f, g with the codomain of f equal to the domain of g, there exists a specified **composite morphism**[4] gf whose domain is equal to the domain of f and whose codomain is equal to the codomain of g, i.e.,:

$$f\colon X \to Y, \quad g\colon Y \to Z \qquad \rightsquigarrow \qquad gf\colon X \to Z\,.$$

This data is subject to the following two axioms:

- For any $f\colon X \to Y$, the composites $1_Y f$ and $f 1_X$ are both equal to f.
- For any composable triple of morphisms f, g, h, the composites $h(gf)$ and $(hg)f$ are equal and henceforth denoted by hgf.

$$f\colon X \to Y, \quad g\colon Y \to Z, \quad h\colon Z \to W \qquad \rightsquigarrow \qquad hgf\colon X \to W\,.$$

That is, the composition law is associative and unital with the identity morphisms serving as two-sided identities.

REMARK 1.1.2. The objects of a category are in bijective correspondence with the identity morphisms, which are uniquely determined by the property that they serve as two-sided identities for composition. Thus, one can define a category to be a collection of morphisms with a partially-defined composition operation that has certain special morphisms, which are used to recognize composable pairs and which serve as two-sided identities; see [**Ehr65**, §I.1] or [**FS90**, §1.1]. But in practice it is not so hard to specify both the objects and the morphisms and this is what we shall do.

It is traditional to name a category after its objects; typically, the preferred choice of accompanying structure-preserving morphisms is clear. However, this practice is somewhat contrary to the basic philosophy of category theory: that mathematical objects should always be considered in tandem with the morphisms between them. By Remark 1.1.2, the algebra of morphisms determines the category, so of the two, the objects and morphisms, the morphisms take primacy.

EXAMPLE 1.1.3. Many familiar varieties of mathematical objects assemble into a category.

[4]The composite may be written less concisely as $g \cdot f$ when this adds typographical clarity.

(i) Set has sets as its objects and functions, with specified domain and codomain,[5] as its morphisms.
(ii) Top has topological spaces as its objects and continuous functions as its morphisms.
(iii) Set_* and Top_* have sets or spaces with a specified basepoint[6] as objects and basepoint-preserving (continuous) functions as morphisms.
(iv) Group has groups as objects and group homomorphisms as morphisms. This example lent the general term "morphisms" to the data of an abstract category. The categories Ring of associative and unital rings and ring homomorphisms and Field of fields and field homomorphisms are defined similarly.
(v) For a fixed unital but not necessarily commutative ring R, Mod_R is the category of left R-modules and R-module homomorphisms. This category is denoted by $\mathsf{Vect}_\Bbbk$ when the ring happens to be a field $\Bbbk$ and abbreviated as Ab in the case of $\mathsf{Mod}_\mathbb{Z}$, as a $\mathbb{Z}$-module is precisely an abelian group.
(vi) Graph has graphs as objects and graph morphisms (functions carrying vertices to vertices and edges to edges, preserving incidence relations) as morphisms. In the variant $\mathsf{DirGraph}$, objects are directed graphs, whose edges are now depicted as arrows, and morphisms are directed graph morphisms, which must preserve sources and targets.
(vii) Man has smooth (i.e., infinitely differentiable) manifolds as objects and smooth maps as morphisms.
(viii) Meas has measurable spaces as objects and measurable functions as morphisms.
(ix) Poset has partially-ordered sets as objects and order-preserving functions as morphisms.
(x) Ch_R has chain complexes of R-modules as objects and chain homomorphisms as morphisms.[7]
(xi) For any *signature* σ, specifying constant, function, and relation symbols, and for any collection of well formed sentences $\mathbb{T}$ in the first-order language associated to σ, there is a category $\mathsf{Model}_\mathbb{T}$ whose objects are σ-structures that *model* $\mathbb{T}$, i.e., sets equipped with appropriate constants, relations, and functions satisfying the axioms $\mathbb{T}$. Morphisms are functions that preserve the specified constants, relations, and functions, in the usual sense.[8] Special cases include (iv), (v), (vi), (ix), and (x).

The preceding are all examples of *concrete categories*, those whose objects have underlying sets and whose morphisms are functions between these underlying sets, typically the "structure-preserving" morphisms. A more precise definition of a concrete category is given in 1.6.17. However, "abstract" categories are also prevalent:

EXAMPLE 1.1.4.

(i) For a unital ring R, Mat_R is the category whose objects are positive integers and in which the set of morphisms from n to m is the set of $m \times n$ matrices with values in

---

[5][**EM45**, p. 239] emphasizes that the data of a function should include specified sets of inputs and potential outputs, a perspective that was somewhat radical at the time.

[6]A **basepoint** is simply a chosen distinguished point in the set or space.

[7]A **chain complex** $C_\bullet$ is a collection $(C_n)_{n\in\mathbb{Z}}$ of R-modules equipped with R-module homomorphisms $d\colon C_n \to C_{n-1}$, called **boundary homomorphisms**, with the property that $d^2 = 0$, i.e., the composite of any two boundary maps is the zero homomorphism. A map of chain complexes $f\colon C_\bullet \to C'_\bullet$ is comprised of a collection of homomorphisms $f_n\colon C_n \to C'_n$ so that $df_n = f_{n-1}d$ for all $n \in \mathbb{Z}$.

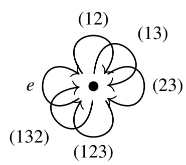
[8]Model theory pays greater attention to other types of morphisms, for instance the *elementary embeddings*, which are (automatically injective) functions that preserve and reflect satisfaction of first-order formulae.

R. Composition is by matrix multiplication

$$n \xrightarrow{A} m, \quad m \xrightarrow{B} k \qquad \rightsquigarrow \qquad n \xrightarrow{B \cdot A} k$$

with identity matrices serving as the identity morphisms.

(ii) A group G (or, more generally, a monoid[9]) defines a category $\mathsf{B}G$ with a single object. The group elements are its morphisms, each group element representing a distinct endomorphism of the single object, with composition given by multiplication. The identity element $e \in G$ acts as the identity morphism for the unique object in this category.

$\mathsf{B}S_3 =$ 

(iii) A poset $(\mathsf{P}, \leq)$ (or, more generally, a preorder[10]) may be regarded as a category. The elements of P are the objects of the category and there exists a unique morphism $x \to y$ if and only if $x \leq y$. Transitivity of the relation "$\leq$" implies that the required composite morphisms exist. Reflexivity implies that identity morphisms exist.

(iv) In particular, any ordinal $\alpha = \{\beta \mid \beta < \alpha\}$ defines a category whose objects are the smaller ordinals. For example, $\mathbb{0}$ is the category with no objects and no morphisms. $\mathbb{1}$ is the category with a single object and only its identity morphism. $\mathbb{2}$ is the category with two objects and a single non-identity morphism, conventionally depicted as $0 \to 1$. ω is the category *freely generated by the graph*

$$0 \to 1 \to 2 \to 3 \to \cdots$$

in the sense that every non-identity morphism can be uniquely factored as a composite of morphisms in the displayed graph; a precise definition of the notion of free generation is given in Example 4.1.13.

(v) A set may be regarded as a category in which the elements of the set define the objects and the only morphisms are the required identities. A category is **discrete** if every morphism is an identity.

(vi) Htpy, like Top, has spaces as its objects but morphisms are homotopy classes of continuous maps. Htpy_* has based spaces as its objects and basepoint-preserving homotopy classes of based continuous maps as its morphisms.

(vii) $\mathsf{Measure}$ has measure spaces as objects. One reasonable choice for the morphisms is to take equivalence classes of measurable functions, where a parallel pair of functions are equivalent if their domain of difference is contained within a set of measure zero.

Thus, the philosophy of category theory is extended. The categories listed in Example 1.1.3 suggest that mathematical objects ought to be considered together with the appropriate notion of morphism between them. The categories listed in Example 1.1.4 illustrate that

[9] A **monoid** is a set M equipped with an associative binary multiplication operation $M \times M \xrightarrow{\cdot} M$ and an identity element $e \in M$ serving as a two-sided identity. In other words, a monoid is precisely a one-object category.

[10] A **preorder** is a set with a binary relation $\leq$ that is reflexive and transitive. In other words, a preorder is precisely a category in which there are no parallel pairs of distinct morphisms between any fixed pair of objects. A **poset** is a preorder that is additionally antisymmetric: $x \leq y$ and $y \leq x$ implies that $x = y$.

these morphisms are not always functions.[11] The morphisms in a category are also called **arrows** or **maps**, particularly in the contexts of Examples 1.1.4 and 1.1.3, respectively.

REMARK 1.1.5. Russell's paradox implies that there is no set whose elements are "all sets." This is the reason why we have used the vague word "collection" in Definition 1.1.1. Indeed, in each of the examples listed in 1.1.3, the collection of objects is not a set. Eilenberg and Mac Lane address this potential area of concern as follows:

> ... the whole concept of a category is essentially an auxiliary one; our basic concepts are essentially those of a *functor* and of a natural transformation .... The idea of a category is required only by the precept that every function should have a definite class as domain and a definite class as range, for the categories are provided as the domains and ranges of functors. Thus one could drop the category concept altogether and adopt an even more intuitive standpoint, in which a functor such as "Hom" is not defined over the category of "all" groups, but for each particular pair of groups which may be given. [**EM45**]

The set-theoretical issues that confront us while defining the notion of a category will compound as we develop category theory further. For that reason, common practice among category theorists is to work in an extension of the usual Zermelo–Fraenkel axioms of set theory, with new axioms allowing one to distinguish between "small" and "large" sets, or between sets and classes. The search for the most useful set-theoretical foundations for category theory is a fascinating topic that unfortunately would require too long of a digression to explore.[12] Instead, we sweep these foundational issues under the rug, not because these issues are not serious or interesting, but because they distract from the task at hand.[13]

For the reasons just discussed, it is important to introduce adjectives that explicitly address the size of a category.

DEFINITION 1.1.6. A category is **small** if it has only a set's worth of arrows.

By Remark 1.1.2, a small category has only a set's worth of objects. If C is a small category, then there are functions

$$\begin{tikzcd} \operatorname{mor}\mathsf{C} \arrow[r, shift left=2, "\mathrm{dom}"] \arrow[r, shift right=2, "\mathrm{cod}"'] & \operatorname{ob}\mathsf{C} \arrow[l, "\mathrm{id}" description] \end{tikzcd}$$

that send a morphism to its domain and its codomain and an object to its identity.

[11] Reid's *Undergraduate algebraic geometry* emphasizes that the morphisms are not always functions, writing "Students who disapprove are recommended to give up at once and take a reading course in category theory instead" [**Rei88**, p. 4].

[12] The preprint [**Shu08**] gives an excellent overview, though it is perhaps better read after Chapters 1–4.

[13] If pressed, let us assume that there exists a countable sequence of **inaccessible cardinals**, meaning uncountable cardinals that are **regular** and **strong limit**. A cardinal κ is **regular** if every union of fewer than κ sets each of cardinality less than κ has cardinality less than κ, and **strong limit** if $\lambda < \kappa$ implies that $2^\lambda < \kappa$. Inaccessibility means that sets of size less than κ are closed under power sets and κ-small unions. If κ is inaccessible, then the κ-stage of the *von Neumann hierarchy*, the set V_κ of sets of rank less than κ, is a model of Zermelo–Fraenkel set theory with choice (ZFC); the set V_κ is a *Grothendieck universe*. The assumption that there exists a countable sequence of inaccessible cardinals means that we can "do set theory" inside the universe V_κ, and then enlarge the universe if necessary as often as needed.

If ZFC is consistent, these axioms cannot prove the existence of an inaccessible cardinal or the consistency of the assumption that one exists (by Gödel's second incompleteness theorem). Nonetheless, from the perspective of the hierarchy of large cardinal axioms, the existence of inaccessibles is a relatively mild hypothesis.

None of the categories in Example 1.1.3 are small—each has too many objects—but "locally" they resemble small categories in a sense made precise by the following notion:

DEFINITION 1.1.7. A category is **locally small** if between any pair of objects there is only a set's worth of morphisms.

It is traditional to write

$$\mathsf{C}(X, Y) \qquad \text{or} \qquad \mathrm{Hom}(X, Y) \tag{1.1.8}$$

for the set of morphisms from X to Y in a locally small category C.[14] The set of arrows between a pair of fixed objects in a locally small category is typically called a **hom-set**, whether or not it is a set of "homomorphisms" of any particular kind. Because the notation (1.1.8) is so convenient, it is also adopted for the collection of morphisms between a fixed pair of objects in a category that is not necessarily locally small.

A category provides a context in which to answer the question "When is one thing the same as another thing?" Almost universally in mathematics, one regards two objects of the same category to be "the same" when they are isomorphic, in a precise categorical sense that we now introduce.

DEFINITION 1.1.9. An **isomorphism** in a category is a morphism $f \colon X \to Y$ for which there exists a morphism $g \colon Y \to X$ so that $gf = 1_X$ and $fg = 1_Y$. The objects X and Y are **isomorphic** whenever there exists an isomorphism between X and Y, in which case one writes $X \cong Y$.

An **endomorphism**, i.e., a morphism whose domain equals its codomain, that is an isomorphism is called an **automorphism**.

EXAMPLE 1.1.10.

(i) The isomorphisms in Set are precisely the **bijections**.
(ii) The isomorphisms in Group, Ring, Field, or Mod_R are the bijective homomorphisms.
(iii) The isomorphisms in the category Top are the **homeomorphisms**, i.e., the continuous functions with continuous inverse, which is a stronger property than merely being a bijective continuous function.
(iv) The isomorphisms in the category Htpy are the **homotopy equivalences**.
(v) In a poset $(\mathsf{P}, \leq)$, the axiom of antisymmetry asserts that $x \leq y$ and $y \leq x$ imply that $x = y$. That is, the only isomorphisms in the category $(\mathsf{P}, \leq)$ are identities.

Examples 1.1.10(ii) and (iii) suggest the following general question: In a concrete category, when are the isomorphisms precisely those maps in the category that induce bijections between the underlying sets? We will see an answer in Lemma 5.6.1.

DEFINITION 1.1.11. A **groupoid** is a category in which every morphism is an isomorphism.

EXAMPLE 1.1.12.

(i) A **group** is a groupoid with one object.[15]
(ii) For any space X, its **fundamental groupoid** $\Pi_1(X)$ is a category whose objects are the points of X and whose morphisms are endpoint-preserving homotopy classes of paths.

---

[14]Mac Lane credits Emmy Noether for emphasizing the importance of homomorphisms in abstract algebra, particularly the homomorphism onto a quotient group, which plays an integral role in the statement of her first isomorphism theorem. His recollection is that the arrow notation first appeared around 1940, perhaps due to Hurewicz [**ML88**]. The notation $\mathrm{Hom}(X, Y)$ was first used in [**EM42a**] for the set of homomorphisms between a pair of abelian groups.

[15]This is not simply an example; it is a definition.

A **subcategory** D of a category C is defined by restricting to a subcollection of objects and subcollection of morphisms subject to the requirements that the subcategory D contains the domain and codomain of any morphism in D, the identity morphism of any object in D, and the composite of any composable pair of morphisms in D. For example, there is a subcategory $\mathsf{CRing} \subset \mathsf{Ring}$ of commutative unital rings. Both of these form subcategories of the category Rng of not-necessarily unital rings and homomorphisms that need not preserve the multiplicative unit.[16]

LEMMA 1.1.13. *Any category C contains a **maximal groupoid**, the subcategory containing all of the objects and only those morphisms that are isomorphisms.*

PROOF. Exercise 1.1.ii. □

For instance, $\mathsf{Fin}_{\mathrm{iso}}$, the category of finite sets and bijections, is the maximal subgroupoid of the category Fin of finite sets and all functions. Example 1.4.9 will explain how this groupoid can be regarded as a categorification of the natural numbers, providing a vantage point from which to prove the laws of elementary arithmetic.

Exercises.

EXERCISE 1.1.i.

(i) Consider a morphism $f\colon x \to y$. Show that if there exists a pair of morphisms $g, h\colon y \rightrightarrows x$ so that $gf = 1_x$ and $fh = 1_y$, then $g = h$ and f is an isomorphism.
(ii) Show that a morphism can have at most one inverse isomorphism.

EXERCISE 1.1.ii. Let C be a category. Show that the collection of isomorphisms in C defines a subcategory, the **maximal groupoid** inside C.

EXERCISE 1.1.iii. For any category C and any object $c \in \mathsf{C}$, show that:

(i) There is a category c/C whose objects are morphisms $f\colon c \to x$ with domain c and in which a morphism from $f\colon c \to x$ to $g\colon c \to y$ is a map $h\colon x \to y$ between the codomains so that the triangle

$$\begin{tikzcd} & c \arrow[dl, "f"'] \arrow[dr, "g"] & \\ x \arrow[rr, "h"'] & & y \end{tikzcd}$$

commutes, i.e., so that $g = hf$.
(ii) There is a category C/c whose objects are morphisms $f\colon x \to c$ with codomain c and in which a morphism from $f\colon x \to c$ to $g\colon y \to c$ is a map $h\colon x \to y$ between the domains so that the triangle

$$\begin{tikzcd} x \arrow[rr, "h"] \arrow[dr, "f"'] & & y \arrow[dl, "g"] \\ & c & \end{tikzcd}$$

commutes, i.e., so that $f = gh$.

The categories c/C and C/c are called **slice categories** of C **under** and **over** c, respectively.

[16] To justify our default notion of ring, see Poonen's "Why all rings should have a 1" [**Poo14**]. The relationship between unital and non-unital rings is explored in greater depth in §4.6.

1.2. Duality

> The dual of any axiom for a category is also an axiom ... A simple metamathematical argument thus proves the *duality principle*. If any statement about a category is deducible from the axioms for a category, the dual statement is likely deducible.
>
> Saunders Mac Lane, "Duality for groups" [**ML50**]

Upon first acquaintance, the primary role played by the notion of a category might appear to be taxonomic: vector spaces and linear maps define one category, manifolds and smooth functions define another. But a category, as defined in 1.1.1, is also a mathematical object in its own right, and as with any mathematical definition, this one is worthy of further consideration. Applying a mathematician's gaze to the definition of a category, the following observation quickly materializes. If we visualize the morphisms in a category as arrows pointing from their domain object to their codomain object, we might imagine simultaneously reversing the directions of every arrow. This leads to the following notion.

DEFINITION 1.2.1. Let C be any category. The **opposite category** C^{op} has

- the same objects as in C, and
- a morphism f^{op} in C^{op} for each a morphism f in C so that the domain of f^{op} is defined to be the codomain of f and the codomain of f^{op} is defined to be the domain of f: i.e.,

$$f^{\mathrm{op}}\colon X \to Y \quad \in \mathsf{C}^{\mathrm{op}} \qquad \leftrightsquigarrow \qquad f\colon Y \to X \quad \in \mathsf{C}\,.$$

That is, C^{op} has the same objects and morphisms as C, except that "each morphism is pointing in the opposite direction." The remaining structure of the category C^{op} is given as follows:

- For each object X, the arrow 1_X^{op} serves as its identity in C^{op}.
- To define composition, observe that a pair of morphisms $f^{\mathrm{op}}, g^{\mathrm{op}}$ in C^{op} is composable precisely when the pair g, f is composable in C, i.e., precisely when the codomain of g equals the domain of f. We then define $g^{\mathrm{op}} \cdot f^{\mathrm{op}}$ to be $(f \cdot g)^{\mathrm{op}}$: i.e.,

$$\begin{array}{ccccc} f^{\mathrm{op}}\colon X \to Y,\ g^{\mathrm{op}}\colon Y \to Z & \in \mathsf{C}^{\mathrm{op}} & \rightsquigarrow & g^{\mathrm{op}} f^{\mathrm{op}}\colon X \to Z & \in \mathsf{C}^{\mathrm{op}} \\ \updownarrow & & & \updownarrow & \\ g\colon Z \to Y,\ f\colon Y \to X & \in \mathsf{C} & \rightsquigarrow & fg\colon Z \to X & \in \mathsf{C} \end{array}$$

The data described in Definition 1.2.1 defines a category C^{op}—i.e., the composition law is associative and unital—if and only if C defines a category. In summary, the process of "turning around the arrows" or "exchanging domains and codomains" exhibits a syntactical self-duality satisfied by the axioms for a category. Note that the category C^{op} contains precisely the same information as the category C. Questions about the one can be answered by examining the other.

EXAMPLE 1.2.2.

(i) $\mathsf{Mat}_R^{\mathrm{op}}$ is the category whose objects are non-zero natural numbers and in which a morphism from m to n is an $m \times n$ matrix with values in R. The upshot is that a reader who would have preferred the opposite handedness conventions when defining Mat_R would have lost nothing by adopting them.

(ii) When a preorder $(\mathsf{P}, \leq)$ is regarded as a category, its opposite category is the category that has a morphism $x \to y$ if and only if $y \leq x$. For example, ω^{op} is the category freely generated by the graph

$$\cdots \to 3 \to 2 \to 1 \to 0\,.$$

(iii) If G is a group, regarded as a one-object groupoid, the category $(\mathsf{B}G)^{\mathrm{op}} \cong \mathsf{B}(G^{\mathrm{op}})$ is again a one-object groupoid, and hence a group. The group G^{op} is called the **opposite group** and is used to define right actions as a special case of left actions; see Example 1.3.9.

This syntactical duality has a very important consequence for the development of category theory. Any theorem containing a universal quantification of the form "for all categories C" also necessarily applies to the opposites of these categories. Interpreting the result in the dual context leads to a **dual theorem**, proven by the dual of the original proof, in which the direction of each arrow appearing in the argument is reversed. The result is a two-for-one deal: any proof in category theory simultaneously proves two theorems, the original statement and its dual.[17] For example, the reader may have found Exercise 1.1.iii redundant, precisely because the statements (i) and (ii) are dual; see Exercise 1.2.i.

To illustrate the principle of duality in category theory, let us consider the following result, which provides an important characterization of the isomorphisms in a category.

LEMMA 1.2.3. *The following are equivalent:*

(i) $f : x \to y$ *is an isomorphism in* C.
(ii) For all objects $c \in \mathsf{C}$*, post-composition with* f *defines a bijection*

$$f_* : \mathsf{C}(c, x) \to \mathsf{C}(c, y)\,.$$

(iii) For all objects $c \in \mathsf{C}$*, pre-composition with* f *defines a bijection*

$$f^* : \mathsf{C}(y, c) \to \mathsf{C}(x, c)\,.$$

REMARK 1.2.4. In language introduced in Chapter 2, Lemma 1.2.3 asserts that isomorphisms in a locally small category are defined *representably* in terms of isomorphisms in the category of sets. That is, a morphism $f : x \to y$ in an arbitrary locally small category C is an isomorphism if and only if the post-composition function $f_* : \mathsf{C}(c, x) \to \mathsf{C}(c, y)$ between hom-sets defines an isomorphism in Set for each object $c \in \mathsf{C}$.

In set theoretical foundations that permit the definition of functions between large sets, the proof given here applies also to non-locally small categories. In our exposition, the set theoretical hypotheses of smallness and local smallness will only appear when there are essential subtleties concerning the sizes of the categories in question. This is not one of those occasions.

PROOF OF LEMMA 1.2.3. We will prove the equivalence (i) $\Leftrightarrow$ (ii) and conclude the equivalence (i) $\Leftrightarrow$ (iii) by duality.

Assuming (i), namely that $f : x \to y$ is an isomorphism with inverse $g : y \to x$, then, as an immediate application of the associativity and identity laws for composition in a category, post-composition with g defines an inverse function

$$g_* : \mathsf{C}(c, y) \to \mathsf{C}(c, x)$$

to f_* in the sense that the composites

$$g_* f_* : \mathsf{C}(c, x) \to \mathsf{C}(c, x) \quad \text{and} \quad f_* g_* : \mathsf{C}(c, y) \to \mathsf{C}(c, y)$$

[17]More generally, the proof of a statement of the form "for all categories $\mathsf{C}_1, \mathsf{C}_2, \ldots, \mathsf{C}_n$" leads to 2^n dual theorems. In practice, however, not all of the dual statements will differ meaningfully from the original; see e.g., §4.3.

are both the identity function: for any $h: c \to x$ and $k: c \to y$, $g_* f_*(h) = gfh = h$ and $f_* g_*(k) = fgk = k$.

Conversely, assuming (ii), there must be an element $g \in \mathsf{C}(y, x)$ whose image under $f_*: \mathsf{C}(y, x) \to \mathsf{C}(y, y)$ is 1_y. By construction, $1_y = fg$. But now, by associativity of composition, the elements $gf, 1_x \in \mathsf{C}(x, x)$ have the common image f under the function $f_*: \mathsf{C}(x, x) \to \mathsf{C}(x, y)$, whence $gf = 1_x$. Thus, f and g are inverse isomorphisms.

We have just proven the equivalence (i) ⇔ (ii) for all categories and in particular for the category C^{op}: i.e., a morphism $f^{\mathrm{op}}: y \to x$ in C^{op} is an isomorphism if and only if

$$f^{\mathrm{op}}_*: \mathsf{C}^{\mathrm{op}}(c, y) \to \mathsf{C}^{\mathrm{op}}(c, x) \text{ is an isomorphism for all } c \in \mathsf{C}^{\mathrm{op}}\,. \tag{1.2.5}$$

Interpreting the data of C^{op} in its opposite category C, the statement (1.2.5) expresses the same mathematical content as

$$f^*: \mathsf{C}(y, c) \to \mathsf{C}(x, c) \text{ is an isomorphism for all } c \in \mathsf{C}\,. \tag{1.2.6}$$

That is: $\mathsf{C}^{\mathrm{op}}(c, x) = \mathsf{C}(x, c)$, post-composition with f^{op} in C^{op} translates to pre-composition with f in the opposite category C. The notion of isomorphism, as defined in 1.1.9, is self-dual: $f^{\mathrm{op}}: y \to x$ is an isomorphism in C^{op} if and only if $f: x \to y$ is an isomorphism in C. So the equivalence (i) ⇔ (ii) in C^{op} expresses the equivalence (i) ⇔ (iii) in C.[18] □

Concise expositions of the duality principle in category theory may be found in [**Awo10**, §3.1] and [**HS97**, §II.3]. As we become more comfortable with arguing by duality, dual proofs and eventually also dual statements will seldom be described in this much detail.

Categorical definitions also have duals; for instance:

Definition 1.2.7. A morphism $f: x \to y$ in a category is

(i) a **monomorphism** if for any parallel morphisms $h, k: w \rightrightarrows x$, $fh = fk$ implies that $h = k$; or
(ii) an **epimorphism** if for any parallel morphisms $h, k: y \rightrightarrows z$, $hf = kf$ implies that $h = k$.

Note that a monomorphism or epimorphism in C is, respectively, an epimorphism or monomorphism in C^{op}. In adjectival form, a monomorphism is **monic** and an epimorphism is **epic**. In common shorthand, a monomorphism is a **mono** and an epimorphism is an **epi**. For graphical emphasis, monos are often decorated with a tail "$\rightarrowtail$" while epis may be decorated at their head "$\twoheadrightarrow$."

The following dual statements re-express Definition 1.2.7:

(i) $f: x \to y$ is a monomorphism in C if and only if for all objects $c \in \mathsf{C}$, post-composition with f defines an injection $f_*: \mathsf{C}(c, x) \to \mathsf{C}(c, y)$.
(ii) $f: x \to y$ is an epimorphism in C if and only if for all objects $c \in \mathsf{C}$, pre-composition with f defines an injection $f^*: \mathsf{C}(y, c) \to \mathsf{C}(x, c)$.

Example 1.2.8. Suppose $f: X \to Y$ is a monomorphism in the category of sets. Then, in particular, given any two maps $x, x': 1 \rightrightarrows X$, whose domain is the singleton set, if $fx = fx'$ then $x = x'$. Thus, monomorphisms are injective functions. Conversely, any injective function can easily be seen to be a monomorphism.

Similarly, a function $f: X \to Y$ is an epimorphism in the category of sets if and only if it is surjective. Given functions $h, k: Y \rightrightarrows Z$, the equation $hf = kf$ says exactly that h is equal to k on the image of f. This only implies that $h = k$ in the case where the image is all of Y.

[18] A similar translation, as just demonstrated between the statements (1.2.5) and (1.2.6), transforms the proof of (i) ⇔ (ii) into a proof of (i) ⇔ (iii).

Thus, monomorphisms and epimorphisms should be regarded as categorical analogs of the notions of injective and surjective functions. In practice, if C is a category in which objects have "underlying sets," then any morphism that induces an injective or surjective function between these defines a monomorphism or epimorphism; see Exercise 1.6.iii for a precise discussion. However, even in such categories, the notions of monomorphism and epimorphism can be more general, as demonstrated in Exercise 1.6.v.

Example 1.2.9. Suppose that $x \xrightarrow{s} y \xrightarrow{r} x$ are morphisms so that $rs = 1_x$. The map s is a **section** or **right inverse** to r, while the map r defines a **retraction** or **left inverse** to s. The maps s and r express the object x as a **retract** of the object y.

In this case, s is always a monomorphism and, dually, r is always an epimorphism. To acknowledge the presence of these one-sided inverses, s is said to be a **split monomorphism** and r is said to be a **split epimorphism**.[19]

Example 1.2.10. By the previous example, an isomorphism is necessarily both monic and epic, but the converse need not hold in general. For example, the inclusion $\mathbb{Z} \hookrightarrow \mathbb{Q}$ is both monic and epic in the category Ring, but this map is not an isomorphism: there are no ring homomorphisms from $\mathbb{Q}$ to $\mathbb{Z}$.

Since the notions of monomorphism and epimorphism are dual, their abstract categorical properties are also dual, such as exhibited by the following lemma.

Lemma 1.2.11.

(i) If $f\colon x \rightarrowtail y$ and $g\colon y \rightarrowtail z$ are monomorphisms, then so is $gf\colon x \rightarrowtail z$.
(ii) If $f\colon x \to y$ and $g\colon y \to z$ are morphisms so that gf is monic, then f is monic.

Dually:

(i') If $f\colon x \twoheadrightarrow y$ and $g\colon y \twoheadrightarrow z$ are epimorphisms, then so is $gf\colon x \twoheadrightarrow z$.
(ii') If $f\colon x \to y$ and $g\colon y \to z$ are morphisms so that gf is epic, then g is epic.

Proof. Exercise 1.2.iii. ◻

Exercises.

Exercise 1.2.i. Show that $\mathsf{C}/c \cong (c/(\mathsf{C}^{\mathrm{op}}))^{\mathrm{op}}$. Defining C/c to be $(c/(\mathsf{C}^{\mathrm{op}}))^{\mathrm{op}}$, deduce Exercise 1.1.iii(ii) from Exercise 1.1.iii(i).

Exercise 1.2.ii.

(i) Show that a morphism $f\colon x \to y$ is a split epimorphism in a category C if and only if for all $c \in \mathsf{C}$, the post-composition function $f_*\colon \mathsf{C}(c, x) \to \mathsf{C}(c, y)$ is surjective.
(ii) Argue by duality that f is a split monomorphism if and only if for all $c \in \mathsf{C}$, the pre-composition function $f^*\colon \mathsf{C}(y, c) \to \mathsf{C}(x, c)$ is surjective.

Exercise 1.2.iii. Prove Lemma 1.2.11 by proving either (i) or (i') and either (ii) or (ii'), then arguing by duality. Conclude that the monomorphisms in any category define a subcategory of that category and dually that the epimorphisms also define a subcategory.

Exercise 1.2.iv. What are the monomorphisms in the category of fields?

Exercise 1.2.v. Show that the inclusion $\mathbb{Z} \hookrightarrow \mathbb{Q}$ is both a monomorphism and an epimorphism in the category Ring of rings. Conclude that a map that is both monic and epic need not be an isomorphism.

Exercise 1.2.vi. Prove that a morphism that is both a monomorphism and a split epimorphism is necessarily an isomorphism. Argue by duality that a split monomorphism that is an epimorphism is also an isomorphism.

[19] The axiom of choice asserts that every epimorphism in the category of sets is a split epimorphism.

EXERCISE 1.2.vii. Regarding a poset $(\mathsf{P}, \leq)$ as a category, define the supremum of a subcollection of objects $A \in \mathsf{P}$ in such a way that the dual statement defines the infimum. Prove that the supremum of a subset of objects is unique, whenever it exists, in such a way that the dual proof demonstrates the uniqueness of the infimum.

1.3. Functoriality

> ... every sufficiently good analogy is yearning to become a functor.
>
> John Baez, "Quantum Quandaries: A Category-Theoretic Perspective" [**Bae06**]

A key tenet in category theory, motivating the very definition of a category, is that any mathematical object should be considered together with its accompanying notion of structure-preserving morphism. In "General theory of natural equivalences" [**EM45**], Eilenberg and Mac Lane argue further:

> ... whenever new abstract objects are constructed in a specified way out of given ones, it is advisable to regard the construction of the corresponding induced mappings on these new objects as an integral part of their definition.

Categories are themselves mathematical objects, if of a somewhat unfamiliar sort, which leads to a question: What is a morphism between categories?

DEFINITION 1.3.1. A **functor** $F \colon \mathsf{C} \to \mathsf{D}$, between categories C and D, consists of the following data:

- An object $Fc \in \mathsf{D}$, for each object $c \in \mathsf{C}$.
- A morphism $Ff \colon Fc \to Fc' \in \mathsf{D}$, for each morphism $f \colon c \to c' \in \mathsf{C}$, so that the domain and codomain of Ff are, respectively, equal to F applied to the domain or codomain of f.

The assignments are required to satisfy the following two **functoriality axioms**:

- For any composable pair f, g in C, $Fg \cdot Ff = F(g \cdot f)$.
- For each object c in C, $F(1_c) = 1_{Fc}$.

Put concisely, a functor consists of a mapping on objects and a mapping on morphisms that preserves all of the structure of a category, namely domains and codomains, composition, and identities.[20]

EXAMPLE 1.3.2.

(i) There is an endofunctor[21] $P \colon \mathsf{Set} \to \mathsf{Set}$ that sends a set A to its power set $PA = \{A' \subset A\}$ and a function $f \colon A \to B$ to the direct-image function $f_* \colon PA \to PB$ that sends $A' \subset A$ to $f(A') \subset B$.

(ii) Each of the categories listed in Example 1.1.3 has a **forgetful functor**, a general term that is used for any functor that forgets structure, whose codomain is the category of sets. For example, $U \colon \mathsf{Group} \to \mathsf{Set}$ sends a group to its underlying set and a group homomorphism to its underlying function. The functor $U \colon \mathsf{Top} \to \mathsf{Set}$ sends a space to its set of points. There are two natural forgetful functors $V, E \colon \mathsf{Graph} \rightrightarrows \mathsf{Set}$ that send a graph to its vertex or edge sets, respectively; if desired, these can be combined

[20]While a functor should be regarded as a mapping from the data of one category to the data of another, parentheses are used as seldom as possible unless demanded for notational clarity.

[21]An **endofunctor** is a functor whose domain is equal to its codomain.

to define a single functor $V \sqcup E\colon \mathsf{Graph} \to \mathsf{Set}$ that carries a graph to the disjoint union of its vertex and edge sets. These mappings are functorial because in each instance a morphism in the domain category has an underlying function.

(iii) There are intermediate forgetful functors $\mathsf{Mod}_R \to \mathsf{Ab}$ and $\mathsf{Ring} \to \mathsf{Ab}$ that forget some but not all of the algebraic structure. The inclusion functors $\mathsf{Ab} \hookrightarrow \mathsf{Group}$ and $\mathsf{Field} \hookrightarrow \mathsf{Ring}$ may also be regarded as "forgetful." Note that the latter two, but neither of the former, are injective on objects: a group is either abelian or not, but an abelian group might admit the structure of a ring in multiple ways.

(iv) Similarly, there are forgetful functors $\mathsf{Group} \to \mathsf{Set}_*$ and $\mathsf{Ring} \to \mathsf{Set}_*$ that take the basepoint to be the identity and zero elements, respectively. These assignments are functorial because group and ring homomorphisms necessarily preserve these elements.

(v) There are functors $\mathsf{Top} \to \mathsf{Htpy}$ and $\mathsf{Top}_* \to \mathsf{Htpy}_*$ that act as the identity on objects and send a (based) continuous function to its homotopy class.

(vi) The *fundamental group* defines a functor $\pi_1\colon \mathsf{Top}_* \to \mathsf{Group}$; a continuous function $f\colon (X, x) \to (Y, y)$ of based spaces induces a group homomorphism $f_*\colon \pi_1(X, x) \to \pi_1(Y, y)$ and this assignment is functorial, satisfying the two functoriality axioms described above. A precise expression of the statement that "the fundamental group is a homotopy invariant" is that this functor factors through the functor $\mathsf{Top}_* \to \mathsf{Htpy}_*$ to define a functor $\pi_1\colon \mathsf{Htpy}_* \to \mathsf{Group}$.

(vii) A related functor $\Pi_1\colon \mathsf{Top} \to \mathsf{Groupoid}$ assigns an unbased topological space its fundamental groupoid, the category defined in Example 1.1.12(ii). A continuous function $f\colon X \to Y$ induces a functor $f_*\colon \Pi_1(X) \to \Pi_1(Y)$ that carries a point $x \in X$ to the point $f(x) \in Y$. This mapping extends to morphisms in $\Pi_1(X)$ because continuous functions preserve paths and path homotopy classes.

(viii) For each $n \in \mathbb{Z}$, there are functors $Z_n, B_n, H_n\colon \mathsf{Ch}_R \to \mathsf{Mod}_R$. The functor Z_n computes the n-**cycles** defined by $Z_nC_\bullet = \ker(d\colon C_n \to C_{n-1})$. The functor B_n computes the n-**boundary** defined by $B_nC_\bullet = \operatorname{im}(d\colon C_{n+1} \to C_n)$. The functor H_n computes the nth **homology** $H_nC_\bullet := Z_nC_\bullet/B_nC_\bullet$. We leave it to the reader to verify that each of these three constructions is functorial. Considering all degrees simultaneously, the cycle, boundary, and homology functors assemble into functors $Z_*, B_*, H_*\colon \mathsf{Ch}_R \to \mathsf{GrMod}_R$ from the category of chain complexes to the category of graded R-modules. The singular homology of a topological space is defined by precomposing H_* with a suitable functor $\mathsf{Top} \to \mathsf{Ch}_R$.

(ix) There is a functor $F\colon \mathsf{Set} \to \mathsf{Group}$ that sends a set X to the **free group** on X. This is the group whose elements are finite "words" whose letters are elements $x \in X$ or their formal inverses x^{-1}, modulo an equivalence relation that equates the words xx^{-1} and $x^{-1}x$ with the empty word. Multiplication is by concatenation, with the empty word serving as the identity. This is one instance of a large family of "free" functors studied in Chapter 4.

(x) The chain rule expresses the functoriality of the derivative. Let Euclid_* denote the category whose objects are pointed finite-dimensional Euclidean spaces $(\mathbb{R}^n, a)$—or, better, open subsets thereof—and whose morphisms are pointed differentiable functions. The **total derivative** of $f\colon \mathbb{R}^n \to \mathbb{R}^m$, evaluated at the designated basepoint $a \in \mathbb{R}^n$, gives rise to a matrix called the **Jacobian matrix** defining the directional derivatives of f at the point a. If f is given by component functions $f_1, \ldots, f_m\colon \mathbb{R}^n \to \mathbb{R}$, the (i, j)-entry of this matrix is $\frac{\partial}{\partial x_j} f_i(a)$. This defines the action on morphisms of a functor $D\colon \mathsf{Euclid}_* \to \mathsf{Mat}_{\mathbb{R}}$; on objects, D assigns a pointed Euclidean space its dimension. Given $g\colon \mathbb{R}^m \to \mathbb{R}^k$ carrying the designated basepoint $f(a) \in \mathbb{R}^m$ to $gf(a) \in \mathbb{R}^k$, functoriality of D asserts that the product of the Jacobian of f at a with the Jacobian of g at $f(a)$ equals the Jacobian of gf at a. This is the chain rule from multivariable calculus.[22]

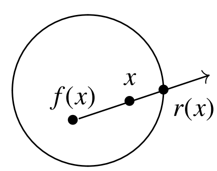
(xi) Any commutative monoid M can be used to define a functor $M^-\colon \mathsf{Fin}_* \to \mathsf{Set}$. Writing $n_+ \in \mathsf{Fin}_*$ for the set with n non-basepoint elements, define M^{n_+} to be M^n, the n-fold cartesian product of the set M with itself. By convention, M^{0_+} is a singleton set. For any based map $f\colon m_+ \to n_+$, define the ith component of the corresponding function $M^f\colon M^m \to M^n$ by projecting from M^m to the coordinates indexed by elements in the fiber $f^{-1}(i)$ and then multiplying these using the commutative monoid structure; if the fiber is empty, the function M^f inserts the unit element in the ith coordinate. Note each of the sets M^n itself has a basepoint, the n-tuple of unit elements, and each of the maps in the image of the functor are based. It follows that the functor M^- lifts along the forgetful functor $U\colon \mathsf{Set}_* \to \mathsf{Set}$.

There is a special property satisfied by this construction that allows one to extract the commutative monoid M from the functor $\mathsf{Fin}_* \to \mathsf{Set}$. This observation was used by Segal to introduce a suitable notion of "commutative monoid" into algebraic topology [**Seg74**].

More examples of functors will appear shortly, but first we illustrate the utility of knowing that the assignment of a mathematical object of one type to mathematical objects of another type is *functorial*. Applying the functoriality of the fundamental group construction $\pi_1\colon \mathsf{Top}_* \to \mathsf{Group}$, one can prove:

Theorem 1.3.3 (Brouwer Fixed Point Theorem). *Any continuous endomorphism of a 2-dimensional disk D^2 has a fixed point.*

Proof. Assuming $f\colon D^2 \to D^2$ is such that $f(x) \neq x$ for all $x \in D^2$, there is a continuous function $r\colon D^2 \to S^1$ that carries a point $x \in D^2$ to the intersection of the ray from $f(x)$ to x with the boundary circle S^1. Note that the function r fixes the points on the boundary circle $S^1 \subset D^2$. Thus, r defines a retraction of the inclusion $i\colon S^1 \hookrightarrow D^2$, which is to say, the composite $S^1 \xrightarrow{i} D^2 \xrightarrow{r} S^1$ is the identity.



Pick any basepoint on the boundary circle S^1 and apply the functor π_1 to obtain a composable pair of group homomorphisms:

$$\pi_1(S^1) \xrightarrow{\pi_1(i)} \pi_1(D^2) \xrightarrow{\pi_1(r)} \pi_1(S^1)\,.$$

By the functoriality axioms, we must have

$$\pi_1(r)\cdot\pi_1(i) = \pi_1(ri) = \pi_1(1_{S^1}) = 1_{\pi_1(S^1)}\,.$$

However, a computation involving covering spaces reveals that $\pi_1(S^1) = \mathbb{Z}$, while $\pi_1(D^2) = 0$, the trivial group. The composite endomorphism $\pi_1(r)\cdot\pi_1(i)$ of $\mathbb{Z}$ must be zero, since it factors through the trivial group. Thus, it cannot equal the identity homomorphism, which

[22]Taking a more sophisticated perspective, the derivative defines the action on morphisms of a functor from the category Man_* to the category of real vector spaces that sends a pointed manifold to its tangent space.

carries the generator $1 \in \mathbb{Z}$ to itself $(0 \neq 1)$. This contradiction proves that the retraction r cannot exist, and so f must have a fixed point.[23] □

Functoriality also plays a key role in the emerging area of topological data analysis.

EXAMPLE 1.3.4 (in search of a clustering functor). A *clustering algorithm* is a function that converts a finite metric space into a partition of its points into sets of "clusters." An impossibility theorem of Kleinberg proves that there are no clustering algorithms that satisfy three reasonable axioms [**Kle03**].[24] A key insight of Carlsson and Mémoli is that these axioms can be encoded as morphisms in a category of finite metric spaces in such a way that what is desired is not a clustering function but a clustering functor into a suitable category [**CM13**]. Ghrist's *Elementary Applied Topology* [**Ghr14**, p. 216] describes this move as follows:

> What is the good of this? Category theory is criticized as an esoteric language: formal and fruitless for conversation. *This is not so.*[25] The virtue of reformulating (the negative) Theorem [of Kleinberg] functorially is a clearer path to a positive statement. If the goal is to have a theory of clustering; if clustering is, properly, a nontrivial functor; if no nontrivial functors between the proposed categories exist; then, naturally, the solution is to alter the domain or codomain categories and classify the ensuing functors. One such modification is to consider a category of persistent clusters.

One pair of categories considered in [**CM13**] are the categories $\mathsf{FinMetric}$, of finite metric spaces and distance non-increasing functions, and $\mathsf{Cluster}$, of clusters and refinements. An object in $\mathsf{Cluster}$ is a partitioned set. Given a function $f : X \to Y$, the preimages of a partition of the set Y define a partition of X. A morphism in $\mathsf{Cluster}$ is a function $f : X \to Y$ of underlying sets so that the given partition on X refines the partition on X defined by the preimages of the given partition on Y.

Carlsson and Mémoli observe that the only scale-invariant functors $\mathsf{FinMetric} \to \mathsf{Cluster}$ either assign each metric space the discrete partition (into singletons) or the indiscrete partition (into a single cluster); both cases fail to satisfy Kleinberg's surjectivity condition. This suggests that clusters should be replaced by a notion of "persistent" clusters. A *persistent cluster* on X is a functor from the poset $([0, \infty), \leq)$ to the poset of clusters on X, where $\phi \leq \psi$ if and only if the partition ϕ refines the partition ψ. The idea is that when the parameter $r \in [0, \infty)$ is small, the partition on X might be very fine, but the clusters are allowed to coalesce as one "zooms out," i.e., as r increases.

There is a category $\mathsf{PCluster}$ whose objects are persistent clusters and whose morphisms are functions of underlying sets $f : X \to Y$ that define morphisms in $\mathsf{Cluster}$ for each $r \in [0, \infty)$. Carlsson and Mémoli prove that there is a unique functor $\mathsf{FinMetric} \to \mathsf{PCluster}$, which takes the metric space with two points of distance r to the persistent cluster with one cluster for $t \geq r$ and two clusters for $0 \leq t < r$ and satisfies two other reasonable conditions; see [**CM13**] for the details.

---

[23]The same argument, with the nth homotopy group functor $\pi_n : \mathsf{Top}_* \to \mathsf{Group}$ in place of π_1, proves that any continuous endomorphism of an n-dimensional disk has a fixed point.

[24]Namely, there are no clustering algorithms that are invariant under rigid scaling, consistent under alterations to the distance function that "sharpen" the point clusters, and have the property that some distance function realizes each possible partition.

[25]Emphasis his.

The functors defined in 1.3.1 are called **covariant** so as to distinguish them from another variety of functor that we now introduce.

DEFINITION 1.3.5. A **contravariant functor** F from C to D is a functor $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$.[26] Explicitly, this consists of the following data:

- An object $Fc \in \mathsf{D}$, for each object $c \in \mathsf{C}$.
- A morphism $Ff\colon Fc' \to Fc \in \mathsf{D}$, for each morphism $f\colon c \to c' \in \mathsf{C}$, so that the domain and codomain of Ff are, respectively, equal to F applied to the codomain or domain of f.

The assignments are required to satisfy the following two **functoriality axioms**:

- For any composable pair f, g in C, $Ff \cdot Fg = F(g \cdot f)$.
- For each object c in C, $F(1_c) = 1_{Fc}$.

NOTATION 1.3.6. To avoid unnatural arrow-theoretic representations, a morphism in the domain of a functor $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$ will always be depicted as an arrow $f\colon c \to c'$ in C, pointing from its domain in C to its codomain in C. Similarly, its image will always be depicted as an arrow $Ff\colon Fc' \to Fc$ in D, pointing from its domain to its codomain. Note that these conventions require that the domain and codomain objects switch their relative places, from left to right, but in examples, for instance in the case where C and D are concrete categories, these positions are the familiar ones. Graphically, the mapping on morphisms given by a contravariant functor is depicted as follows:

$$\mathsf{C}^{\mathrm{op}} \xrightarrow{F} \mathsf{D}$$

$$\begin{tikzcd} c \arrow[d, "f"'] & \mapsto & Fc \\ {} & \mapsto & {} \\ c' & \mapsto & Fc' \arrow[u, "Ff"'] \end{tikzcd}$$

In accordance with this convention, if $f\colon c \to c'$ and $g\colon c' \to c''$ are morphisms in C, their composite will always be written as $gf\colon c \to c''$. The image of this morphism under the contravariant functor $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$ is $F(gf)\colon Fc'' \to Fc$, the composite $Ff \cdot Fg$ of $Fg\colon Fc'' \to Fc'$ and $Ff\colon Fc' \to Fc$.

In summary, even in the presence of opposite categories, we always make an effort to draw arrows pointing in the "correct way" and depict composition in the usual order.[27]

EXAMPLE 1.3.7.

(i) The contravariant power set functor $P\colon \mathsf{Set}^{\mathrm{op}} \to \mathsf{Set}$ sends a set A to its power set PA and a function $f\colon A \to B$ to the inverse-image function $f^{-1}\colon PB \to PA$ that sends $B' \subset B$ to $f^{-1}(B') \subset A$.

(ii) There is a functor $(-)^*\colon \mathsf{Vect}_{\Bbbk}^{\mathrm{op}} \to \mathsf{Vect}_{\Bbbk}$ that carries a vector space to its **dual vector space** $V^* = \mathrm{Hom}(V, \Bbbk)$. A vector in V^* is a **linear functional** on V, i.e., a linear map

[26] In this text, a contravariant functor F from C to D will always be written as $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$. Some mathematicians omit the "op" and let the context or surrounding verbiage convey the variance. We think this is bad practice, as the co- or contravariance is an essential part of the data of a functor, which is not necessarily determined by its assignation on objects. More to the point, we find that this notational convention helps mitigate the consequences of temporary distraction. Seeing $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$ written on a chalkboard immediately conveys that F is a contravariant functor from C to D, even to the most spaced-out observer. A similar principle will motivate other notational conventions introduced in Definition 3.1.15 and Notation 4.1.5.

[27] Of course, technically there is no meaning to the phrase "opposite category": every category is the opposite of some other category (its opposite category). But in practice, there is no question which of Set and $\mathsf{Set}^{\mathrm{op}}$ is the "opposite category," and sufficiently many of the other cases can be deduced from this one.

$V \to \Bbbk$. This functor is contravariant, with a linear map $\phi \colon V \to W$ sent to the linear map $\phi^* \colon W^* \to V^*$ that pre-composes a linear functional $W \xrightarrow{\omega} \Bbbk$ with ϕ to obtain a linear functional $V \xrightarrow{\phi} W \xrightarrow{\omega} \Bbbk$.

(iii) The functor $\mathcal{O} \colon \mathsf{Top}^{\mathrm{op}} \to \mathsf{Poset}$ that carries a space X to its poset $\mathcal{O}(X)$ of open subsets is contravariant on the category of spaces: a continuous map $f \colon X \to Y$ gives rise to a function $f^{-1} \colon \mathcal{O}(Y) \to \mathcal{O}(X)$ that carries an open subset $U \subset Y$ to its preimage $f^{-1}(U)$, which is open in X; this is the definition of **continuity**. A similar functor $\mathcal{C} \colon \mathsf{Top}^{\mathrm{op}} \to \mathsf{Poset}$ carries a space to its poset of closed subsets.

(iv) There is a contravariant functor $\mathrm{Spec} \colon \mathsf{CRing}^{\mathrm{op}} \to \mathsf{Top}$ that sends a commutative ring R to its set $\mathrm{Spec}(R)$ of prime ideals given the **Zariski topology**. The closed subsets in the Zariski topology are those subsets $V_I \subset \mathrm{Spec}(R)$ of prime ideals containing a fixed ideal $I \subset R$. This construction is contravariantly functorial: for any ring homomorphism $\phi \colon R \to S$ and prime ideal $\mathfrak{p} \subset S$, the inverse image $\phi^{-1}(\mathfrak{p}) \subset R$ defines a prime ideal of R, and the inverse image function $\phi^{-1} \colon \mathrm{Spec}(S) \to \mathrm{Spec}(R)$ is continuous with respect to the Zariski topology.

(v) For a generic small category C, a functor $\mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ is called a (set-valued) **presheaf** on C. A typical example is the functor $\mathcal{O}(X)^{\mathrm{op}} \to \mathsf{Set}$ whose domain is the poset $\mathcal{O}(X)$ of open subsets of a topological space X and whose value at $U \subset X$ is the set of continuous real-valued functions on U. The action on morphisms is by restriction. This presheaf is a *sheaf*, if it satisfies an axiom that is introduced in Definition 3.3.4.

(vi) Presheaves on the category $\mathbb{\Delta}$, of finite non-empty ordinals and order-preserving maps, are called **simplicial sets**. $\mathbb{\Delta}$ is also called the **simplex category**. The ordinal $n + 1 = \{0, 1, \ldots, n\}$ may be thought of as a direct version of the topological n-simplex and, with this interpretation in mind, is typically denoted by "$[n]$" by algebraic topologists.

The following result, which appears immediately after functors are first defined in [**EM42b**], is arguably the first lemma in category theory.

LEMMA 1.3.8. *Functors preserve isomorphisms.*

PROOF. Consider a functor $F \colon \mathsf{C} \to \mathsf{D}$ and an isomorphism $f \colon x \to y$ in C with inverse $g \colon y \to x$. Applying the two functoriality axioms:

$$F(g)F(f) = F(gf) = F(1_x) = 1_{Fx}.$$

Thus, $Fg \colon Fy \to Fx$ is a left inverse to $Ff \colon Fx \to Fy$. Exchanging the roles of f and g (or arguing by duality) shows that Fg is also a right inverse. □

EXAMPLE 1.3.9. Let G be a group, regarded as a one-object category $\mathsf{B}G$. A functor $X \colon \mathsf{B}G \to \mathsf{C}$ specifies an object $X \in \mathsf{C}$ (the unique object in its image) together with an endomorphism $g_* \colon X \to X$ for each $g \in G$. This assignment must satisfy two conditions:

(i) $h_* g_* = (hg)_*$ for all $g, h \in G$.
(ii) $e_* = 1_X$, where $e \in G$ is the identity element.

In summary, the functor $\mathsf{B}G \to \mathsf{C}$ defines an **action** of the group G on the object $X \in \mathsf{C}$. When $\mathsf{C} = \mathsf{Set}$, the object X endowed with such an action is called a **G-set**. When $\mathsf{C} = \mathsf{Vect}_{\Bbbk}$, the object X is called a **G-representation**. When $\mathsf{C} = \mathsf{Top}$, the object X is called a **G-space**. Note the utility of this categorical language for defining several analogous concepts simultaneously.

The action specified by a functor $\mathsf{B}G \to \mathsf{C}$ is sometimes called a **left action**. A **right action** is a functor $\mathsf{B}G^{\mathrm{op}} \to \mathsf{C}$. As before, each $g \in G$ determines an endomorphism

$g^*\colon X \to X$ in C and the identity element must act trivially. But now, for a pair of elements $g, h \in G$ these actions must satisfy the composition rule $(hg)^* = g^*h^*$.

Because the elements $g \in G$ are isomorphisms when regarded as morphisms in the 1-object category $\mathsf{B}G$ that represents the group, their images under any such functor must also be isomorphisms in the target category. In particular, in the case of a G-representation $V\colon \mathsf{B}G \to \mathsf{Vect}_{\Bbbk}$, the linear map $g_*\colon V \to V$ must be an *automorphism* of the vector space V. The point is that the functoriality axioms (i) and (ii) imply automatically that each g_* is an automorphism and that $(g^{-1})_* = (g_*)^{-1}$; the proof is a special case of Lemma 1.3.8.

In summary:

Corollary 1.3.10. *When a group G acts functorially on an object X in a category C, its elements g must act by automorphisms $g_*\colon X \to X$ and, moreover, $(g_*)^{-1} = (g^{-1})_*$.*

A functor may or may not preserve monomorphisms or epimorphisms, but an argument similar to the proof of Lemma 1.3.8 shows that a functor necessarily preserves split monomorphisms and split epimorphisms. The retraction or section defines an "equational witness" for the mono or the epi.

Definition 1.3.11. If C is locally small, then for any object $c \in \mathsf{C}$ we may define a pair of covariant and contravariant **functors represented by** c:

$$\begin{array}{ccc} \mathsf{C} & \xrightarrow{\mathsf{C}(c,-)} & \mathsf{Set} \\ \\ x & \mapsto & \mathsf{C}(c,x) \\ {\scriptstyle f}\downarrow & \mapsto & \downarrow{\scriptstyle f_*} \\ y & \mapsto & \mathsf{C}(c,y) \end{array} \qquad \begin{array}{ccc} \mathsf{C}^{\mathrm{op}} & \xrightarrow{\mathsf{C}(-,c)} & \mathsf{Set} \\ \\ x & \mapsto & \mathsf{C}(x,c) \\ {\scriptstyle f}\downarrow & \mapsto & \uparrow{\scriptstyle f^*} \\ y & \mapsto & \mathsf{C}(y,c) \end{array}$$

The notation suggests the action on objects: the functor $\mathsf{C}(c,-)$ carries $x \in \mathsf{C}$ to the set $\mathsf{C}(c,x)$ of arrows from c to x in C. Dually, the functor $\mathsf{C}(-,c)$ carries $x \in \mathsf{C}$ to the set $\mathsf{C}(x,c)$.

The functor $\mathsf{C}(c,-)$ carries a morphism $f\colon x \to y$ to the post-composition function $f_*\colon \mathsf{C}(c,x) \to \mathsf{C}(c,y)$ introduced in Lemma 1.2.3(ii). Dually, the functor $\mathsf{C}(-,c)$ carries f to the pre-composition function $f^*\colon \mathsf{C}(y,c) \to \mathsf{C}(x,c)$ introduced in 1.2.3(iii). Note that post-composition defines a *covariant* action on hom-sets, while pre-composition defines a *contravariant* action. There are no choices involved here; post-composition is always a covariant operation, while pre-composition is always a contravariant one. This is just the natural order of things.

We leave it to the reader to verify that the assignments just described satisfy the two functoriality axioms. Note that Lemma 1.3.8 specializes in the case of represented functors to give a proof of the implications (i) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (iii) of Lemma 1.2.3. These functors will play a starring role in Chapter 2, where a number of examples in disguise are discussed.

The data of the covariant and contravariant functors introduced in Definition 1.3.11 may be encoded in a single **bifunctor**, which is the name for a functor of two variables. Its domain is given by the product of a pair of categories.

Definition 1.3.12. For any categories C and D, there is a category $\mathsf{C} \times \mathsf{D}$, their **product**, whose

- objects are ordered pairs (c, d), where c is an object of C and d is an object of D,
- morphisms are ordered pairs $(f, g)\colon (c, d) \to (c', d')$, where $f\colon c \to c' \in \mathsf{C}$ and $g\colon d \to d' \in \mathsf{D}$, and

- in which composition and identities are defined componentwise.

DEFINITION 1.3.13. If C is locally small, then there is a **two-sided represented functor**

$$\mathsf{C}(-,-)\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{C} \to \mathsf{Set}$$

defined in the evident manner. A pair of objects (x, y) is mapped to the hom-set $\mathsf{C}(x, y)$. A pair of morphisms $f\colon w \to x$ and $h\colon y \to z$ is sent to the function

$$\begin{array}{ccc} \mathsf{C}(x,y) & \xrightarrow{(f^*,h_*)} & \mathsf{C}(w,z) \\ g & \mapsto & hgf \end{array}$$

that takes an arrow $g\colon x \to y$ and then pre-composes with f and post-composes with h to obtain $hgf\colon w \to z$.

At the beginning of this section, it was suggested that functors define morphisms between categories. Indeed, categories and functors assemble into a category. Here the size issues are even more significant than we have encountered thus far. To put a lid on things, define Cat to be the category whose objects are small categories and whose morphisms are functors between them. This category is locally small but not small: it contains Set, Poset, Monoid, Group, and $\mathsf{Groupoid}$ as proper subcategories (see Exercises 1.3.i and 1.3.ii). However, none of these categories are *objects* of Cat.

The non-small categories of Example 1.1.3 are objects of CAT, some category of "large" categories and functors between them. Russell's paradox suggests that CAT should not be so large as to contain itself, so we require the objects in CAT to be locally small categories; the category CAT defined in this way is not locally small, and so is thus excluded. There is an inclusion functor $\mathsf{Cat} \hookrightarrow \mathsf{CAT}$ but no obvious functor pointing in the other direction.

The category of categories gives rise to a notion of an **isomorphism of categories**, defined by interpreting Definition 1.1.9 in Cat or in CAT. Namely, an isomorphism of categories is given by a pair of inverse functors $F\colon \mathsf{C} \to \mathsf{D}$ and $G\colon \mathsf{D} \to \mathsf{C}$ so that the composites GF and FG, respectively, equal the identity functors on C and on D. An isomorphism induces a bijection between the objects of C and objects of D and likewise for the morphisms.

EXAMPLE 1.3.14. For instance:

(i) The functor $(-)^{\mathrm{op}}\colon \mathsf{CAT} \to \mathsf{CAT}$ defines a non-trivial automorphism of the category of categories. Note that a functor $F\colon \mathsf{C} \to \mathsf{D}$ also defines a functor $F\colon \mathsf{C}^{\mathrm{op}} \to \mathsf{D}^{\mathrm{op}}$.

(ii) For any group G, the categories $\mathsf{B}G$ and $\mathsf{B}G^{\mathrm{op}}$ are isomorphic via the functor $(-)^{-1}$ that sends each morphism $g \in G$ to its inverse. Any right action can be converted into a left action by precomposing with this isomorphism, which has the effect of "inserting inverses in the formula" defining the endomorphism associated to a particular group element.

(iii) Similarly, any groupoid is isomorphic to its opposite category via the functor that acts as the identity on objects and sends a morphism to its unique inverse morphism.

(iv) Any ring R has an opposite ring R^{op} with the same underlying abelian group but with the product of elements r and s in R^{op} defined to be the product $s \cdot r$ of the elements s and r in R. A left R-module is the same thing as a right R^{op}-module, which is to say there is a covariant isomorphism of categories $\mathsf{Mod}_R \cong {}_{R^{\mathrm{op}}}\mathsf{Mod}$ between the category of left R-modules and the category of right R^{op}-modules.

(v) For any space X, there is a contravariant isomorphism of poset categories $\mathcal{O}(X) \cong \mathcal{C}(X)^{\mathrm{op}}$ that associates an open subset of X to its closed complement.

(vi) The category Mat_R is isomorphic to its opposite via an identity-on-objects functor that carries a matrix to its transpose.

Contrary to the impression created by Examples 1.3.14 (ii), (iii), and (vi), a category is not typically isomorphic to its opposite category.

EXAMPLE 1.3.15. Let E/F be a finite **Galois extension**: this means that F is a finite-index subfield of E and that the size of the group $\mathrm{Aut}(E/F)$ of automorphisms of E fixing every element of F is at least (in fact, equal to) the index $[E : F]$. In this case, $G := \mathrm{Aut}(E/F)$ is called the **Galois group** of the Galois extension E/F.

Consider the **orbit category** $\mathcal{O}_G$ associated to the group G. Its objects are subgroups $H \subset G$, which we identify with the left G-set G/H of left cosets of H. Morphisms $G/H \to G/K$ are G-equivariant maps, i.e., functions that commute with the left G-action. By an elementary exercise left to the reader, every morphism $G/H \to G/K$ has the form $gH \mapsto g\gamma K$, where $\gamma \in G$ is an element so that $\gamma^{-1}H\gamma \subset K$.

Let Field_F^E denote the subcategory of F/Field whose objects are intermediate fields $F \subset K \subset E$. A morphism $K \to L$ is a field homomorphism that fixes the elements of F pointwise. Note that the group of automorphisms of the object $E \in \mathsf{Field}_F^E$ is the Galois group $G = \mathrm{Aut}(E/F)$.

We define a functor $\Phi\colon \mathcal{O}_G^{\mathrm{op}} \to \mathsf{Field}_F^E$ that sends $H \subset G$ to the subfield of E of elements that are fixed by H under the action of the Galois group. If $G/H \to G/K$ is induced by γ, then the field homomorphism $x \mapsto \gamma x$ sends an element $x \in E$ that is fixed by K to an element $\gamma x \in E$ that is fixed by H. This defines the action of the functor Φ on morphisms. The **fundamental theorem of Galois theory** asserts that Φ defines a bijection on objects but in fact more is true: Φ defines an isomorphism of categories $\mathcal{O}_G^{\mathrm{op}} \cong \mathsf{Field}_F^E$.

These examples aside, the notion of isomorphism of categories is somewhat unnatural. To illustrate, consider the category Set^∂ of sets and partially-defined functions. A **partial function** $f\colon X \to Y$ is a function from a (possibly-empty) subset $X' \subset X$ to Y; the subset X' is the **domain of definition** of the partial function f. The composite of two partial functions $f\colon X \to Y$ and $g\colon Y \to Z$ is the partial function whose domain of definition is the intersection of the domain of definition of f with the preimage of the domain of definition of g.

There is a functor $(-)_+\colon \mathsf{Set}^\partial \to \mathsf{Set}_*$, whose codomain is the category of pointed sets, that sends a set X to the pointed set X_+, which is defined to be the disjoint union of X with a freely-added basepoint. By the axiom of regularity, we might define $X_+ := X \cup \{X\}$.[28] A partial function $f\colon X \to Y$ gives rise to a pointed function $f_+\colon X_+ \to Y_+$ that sends every point outside of the domain of definition of f to the formally added basepoint of Y_+. The inverse functor $U\colon \mathsf{Set}_* \to \mathsf{Set}^\partial$ discards the basepoint and sends a based function $f\colon (X, x) \to (Y, y)$ to the partial function $X\backslash\{x\} \to Y\backslash\{y\}$ with the maximal possible domain of definition.

By construction, we see that the composite $U(-)_+$ is the identity endofunctor of the category Set^∂. By contrast, the other composite $(U-)_+\colon \mathsf{Set}_* \to \mathsf{Set}_*$ sends a pointed set (X, x) to $(X\backslash\{x\} \cup \{X\backslash\{x\}\}, X\backslash\{x\})$. These sets are isomorphic but they are not identical. Nor is another set-theoretical construction of the "freely added basepoint" likely to define a genuine inverse to the functor $U\colon \mathsf{Set}_* \to \mathsf{Set}^\partial$. It is too restrictive to ask for the categories Set^∂ and Set_* to be isomorphic.

---

[28]In the axioms of Zermelo-Fraenkel set theory, elements of sets (like everything else in its mathematical universe) are themselves sets. The axiom of regularity prohibits a set from being an element of itself. As $X \notin X$, we are free to add the element X as a disjoint basepoint.

Indeed, there is a better way to decide whether two categories may safely be regarded as "the same." To define it, we must relax the identities $GF = 1_{\mathsf{C}}$ and $FG = 1_{\mathsf{D}}$ between functors $F\colon \mathsf{C} \to \mathsf{D}$ and $G\colon \mathsf{D} \to \mathsf{C}$ that define an isomorphism of categories. This is possible because the collections $\mathrm{Hom}(\mathsf{C}, \mathsf{C})$ and $\mathrm{Hom}(\mathsf{D}, \mathsf{D})$ are not mere (possibly large) sets: they have higher-dimensional structure. For any pair of categories C and D, the collection $\mathrm{Hom}(\mathsf{C}, \mathsf{D})$ of functors is itself a category. To explain this, we introduce what in French is called a *morphisme de foncteurs*, the notion that launched the entire subject of category theory: a *natural transformation*.

Exercises.

EXERCISE 1.3.i. What is a functor between groups, regarded as one-object categories?

EXERCISE 1.3.ii. What is a functor between preorders, regarded as categories?

EXERCISE 1.3.iii. Find an example to show that the objects and morphisms in the image of a functor $F\colon \mathsf{C} \to \mathsf{D}$ do not necessarily define a subcategory of D.

EXERCISE 1.3.iv. Verify that the constructions introduced in Definition 1.3.11 are functorial.

EXERCISE 1.3.v. What is the difference between a functor $\mathsf{C}^{\mathrm{op}} \to \mathsf{D}$ and a functor $\mathsf{C} \to \mathsf{D}^{\mathrm{op}}$? What is the difference between a functor $\mathsf{C} \to \mathsf{D}$ and a functor $\mathsf{C}^{\mathrm{op}} \to \mathsf{D}^{\mathrm{op}}$?

EXERCISE 1.3.vi. Given functors $F\colon \mathsf{D} \to \mathsf{C}$ and $G\colon \mathsf{E} \to \mathsf{C}$, show that there is a category, called the **comma category** $F \downarrow G$, which has

- as objects, triples $(d \in \mathsf{D}, e \in \mathsf{E}, f\colon Fd \to Ge \in \mathsf{C})$, and
- as morphisms $(d, e, f) \to (d', e', f')$, a pair of morphisms $(h\colon d \to d', k\colon e \to e')$ so that the square

$$\begin{tikzcd} Fd \arrow[r, "f"] \arrow[d, "Fh"'] & Ge \arrow[d, "Gk"] \\ Fd' \arrow[r, "f'"'] & Ge' \end{tikzcd}$$

  commutes in C, i.e., so that $f' \cdot Fh = Gk \cdot f$.

Define a pair of projection functors $\mathrm{dom}\colon F \downarrow G \to \mathsf{D}$ and $\mathrm{cod}\colon F \downarrow G \to \mathsf{E}$.

EXERCISE 1.3.vii. Define functors to construct the slice categories c/C and C/c of Exercise 1.1.iii as special cases of comma categories constructed in Exercise 1.3.vi. What are the projection functors?

EXERCISE 1.3.viii. Lemma 1.3.8 shows that functors preserve isomorphisms. Find an example to demonstrate that functors need not **reflect isomorphisms**: that is, find a functor $F\colon \mathsf{C} \to \mathsf{D}$ and a morphism f in C so that Ff is an isomorphism in D but f is not an isomorphism in C.

EXERCISE 1.3.ix. For any group G, we may define other groups:

- the **center** $Z(G) = \{h \in G \mid hg = gh \,\forall g \in G\}$, a subgroup of G,
- the **commutator subgroup** $C(G)$, the subgroup of G generated by elements $ghg^{-1}h^{-1}$ for any $g, h \in G$, and
- the **automorphism group** $\mathrm{Aut}(G)$, the group of isomorphisms $\phi\colon G \to G$ in Group.

Trivially, all three constructions define a functor from the discrete category of groups (with only identity morphisms) to Group. Are these constructions functorial in

- the isomorphisms of groups? That is, do they extend to functors $\mathsf{Group}_{\mathrm{iso}} \to \mathsf{Group}$?

- the epimorphisms of groups[29]? That is, do they extend to functors $\mathsf{Group}_{\mathrm{epi}} \to \mathsf{Group}$?
- all homomorphisms of groups? That is, do they extend to functors $\mathsf{Group} \to \mathsf{Group}$?

Exercise 1.3.x. Show that the construction of the set of conjugacy classes of elements of a group is functorial, defining a functor $\mathrm{Conj}\colon \mathsf{Group} \to \mathsf{Set}$. Conclude that any pair of groups whose sets of conjugacy classes of elements have differing cardinalities cannot be isomorphic.

1.4. Naturality

> It is not too misleading, at least historically, to say that categories are what one must define in order to define functors, and that functors are what one must define in order to define natural transformations.
>
> Peter Freyd, *Abelian categories* [**Fre03**]

Any finite-dimensional $\Bbbk$-vector space V is isomorphic to its **linear dual**, the vector space $V^* = \mathrm{Hom}(V, \Bbbk)$ of linear maps $V \to \Bbbk$, because these vector spaces have the same dimension. This can be proven through the construction of an explicit **dual basis**: choose a basis $e_1, \ldots, e_n$ for V and then define $e_1^*, \ldots, e_n^* \in V^*$ by

$$e_i^*(e_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

The collection $e_1^*, \ldots, e_n^*$ defines a basis for V^* and the map $e_i \mapsto e_i^*$ extends by linearity to define an isomorphism $V \cong V^*$.

Now consider a related construction of the **double dual** $V^{**} = \mathrm{Hom}(\mathrm{Hom}(V, \Bbbk), \Bbbk)$ of V. If V is finite dimensional, then the isomorphism $V \cong V^*$ is carried by the dual vector space functor $(-)^*\colon \mathsf{Vect}_{\Bbbk}^{\mathrm{op}} \to \mathsf{Vect}_{\Bbbk}$ to an isomorphism $V^* \cong V^{**}$. The composite isomorphism $V \cong V^{**}$ sends the basis $e_1, \ldots, e_n$ to the dual dual basis $e_1^{**}, \ldots, e_n^{**}$.

As it turns out, this isomorphism has a simpler description. For any $v \in V$, the "evaluation function"

$$f \mapsto f(v)\colon V^* \xrightarrow{\mathrm{ev}_v} \Bbbk$$

defines a linear functional on V^*. It turns out the assignment $v \mapsto \mathrm{ev}_v$ defines a linear isomorphism $V \cong V^{**}$, this time requiring no "unnatural" choice of basis.[30]

What distinguishes the isomorphism between a finite-dimensional vector space and its double dual from the isomorphism between a finite-dimensional vector space and its single dual is that the former assembles into the components of a *natural transformation* in the sense that we now introduce.

Definition 1.4.1. Given categories C and D and functors $F, G\colon \mathsf{C} \rightrightarrows \mathsf{D}$, a **natural transformation** $\alpha\colon F \Rightarrow G$ consists of:

- an arrow $\alpha_c\colon Fc \to Gc$ in D for each object $c \in \mathsf{C}$, the collection of which define the **components** of the natural transformation,

[29]A non-trivial theorem demonstrates that a homomorphism $\phi\colon G \to H$ is an epimorphism in Group if and only if its underlying function is surjective; see [**Lin70**].

[30]In fact, $e_i^{**}(e_j^*) = e_j^*(e_i) = \mathrm{ev}_{e_i}(e_j^*)$, and so the two isomorphisms $V \cong V^{**}$ are the same—it is only our description that has improved.

so that, for any morphism $f\colon c \to c'$ in C, the following square of morphisms in D

$$\begin{tikzcd} Fc \arrow[r, "\alpha_c"] \arrow[d, "Ff"'] & Gc \arrow[d, "Gf"] \\ Fc' \arrow[r, "\alpha_{c'}"'] & Gc' \end{tikzcd} \tag{1.4.2}$$

commutes, i.e., has a common composite $Fc \to Gc'$ in D.

A **natural isomorphism** is a natural transformation $\alpha\colon F \Rightarrow G$ in which every component α_c is an isomorphism. In this case, the natural isomorphism may be depicted as $\alpha\colon F \cong G$.

In practice, it is usually most elegant to define a natural transformation by saying that "the arrows X are natural," which means that the collection of arrows defines the components of a natural transformation, leaving implicit the correct choices of domain and codomain functors, and source and target categories. Here X should be a collection of morphisms in a clearly identifiable (target) category, whose domains and codomains are defined using a common "variable" (an object of the source category). If this variable is c, one might say "the arrows X are natural in c" to emphasize the domain object whose component is being described. However, the totality of the data of the source and target categories, the parallel pair of functors, and the components should always be considered part of the natural transformation. The naturality condition (1.4.2) cannot be stated precisely with any less: it refers to every object and every morphism in the domain category and is described using the images in the codomain category under the action of both functors. The "boundary data" needed to define a natural transformation α is often displayed in a globular diagram:

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left, "F"{name=U}] \arrow[r, bend right, "G"'{name=D}] & \mathsf{D} \arrow[from=U, to=D, phantom, "\Downarrow\alpha"] \end{tikzcd}$$

The globular depiction of a natural transformation makes the notions of composable natural transformations that are introduced in §1.7 particularly intuitive.

EXAMPLE 1.4.3.

(i) For vector spaces of any dimension, the map $\mathrm{ev}\colon V \to V^{**}$ that sends $v \in V$ to the linear function $\mathrm{ev}_v\colon V^* \to \Bbbk$ defines the components of a natural transformation from the identity endofunctor on $\mathsf{Vect}_\Bbbk$ to the double dual functor. To check that the naturality square

$$\begin{tikzcd} V \arrow[r, "\mathrm{ev}"] \arrow[d, "\phi"'] & V^{**} \arrow[d, "\phi^{**}"] \\ W \arrow[r, "\mathrm{ev}"] & W^{**} \end{tikzcd}$$

commutes for any linear map $\phi\colon V \to W$, it suffices to consider the image of a generic vector $v \in V$. By definition, $\mathrm{ev}_{\phi v}\colon W^* \to \Bbbk$ carries a functional $f\colon W \to \Bbbk$ to $f(\phi v)$. Recalling the definition of the action of the dual functor of Example 1.3.7(ii) on morphisms, we see that $\phi^{**}(\mathrm{ev}_v)\colon W^* \to \Bbbk$ carries a functional $f\colon W \to \Bbbk$ to $f\phi(v)$, which amounts to the same thing.

(ii) By contrast, the identity functor and the single dual functor on finite-dimensional vector spaces are not naturally isomorphic. One technical obstruction is somewhat beside the point: the identity functor is covariant while the dual functor is contravariant.[31] More significant is the essential failure of naturality. The isomorphisms $V \cong V^*$ that can be defined whenever V is finite dimensional require a choice of basis, which is preserved by essentially no linear maps, indeed by no non-identity linear endomorphism.[32]

(iii) There is a natural transformation $\eta \colon 1_{\mathsf{Set}} \Rightarrow P$ from the identity to the covariant power set functor whose components $\eta_A \colon A \to PA$ are the functions that carry $a \in A$ to the singleton subset $\{a\} \in PA$.

(iv) For G a group, Example 1.3.9 shows that a functor $X \colon \mathsf{B}G \to \mathsf{C}$ corresponds to an object $X \in \mathsf{C}$ equipped with a left action of G, which suggests a question: What is a natural transformation between a pair $X, Y \colon \mathsf{B}G \rightrightarrows \mathsf{C}$ of such functors? Because the category $\mathsf{B}G$ has only one object, the data of $\alpha \colon X \Rightarrow Y$ consists of a single morphism $\alpha \colon X \to Y$ in C that is G**-equivariant**, meaning that for each $g \in G$, the diagram

$$\begin{tikzcd} X \arrow[r, "\alpha"] \arrow[d, "g_*"'] & Y \arrow[d, "g_*"] \\ X \arrow[r, "\alpha"'] & Y \end{tikzcd}$$

commutes.

(v) The open and closed subset functors described in Example 1.3.7(iii) are naturally isomorphic when they are regarded as functors $O, C \colon \mathsf{Top}^{\mathrm{op}} \rightrightarrows \mathsf{Set}$ valued in the category of sets. The components $O(X) \cong C(X)$ of the natural isomorphism are defined by taking an open subset of X to its complement, which is closed. Naturality asserts that the process of forming complements commutes with the operation of taking preimages.

(vi) The construction of the opposite group described in Example 1.2.2(iii) defines a (covariant!) endofunctor $(-)^{\mathrm{op}} \colon \mathsf{Group} \to \mathsf{Group}$ of the category of groups; a homomorphism $\phi \colon G \to H$ induces a homomorphism $\phi^{\mathrm{op}} \colon G^{\mathrm{op}} \to H^{\mathrm{op}}$ defined by $\phi^{\mathrm{op}}(g) = \phi(g)$. This functor is naturally isomorphic to the identity. Define $\eta_G \colon G \to G^{\mathrm{op}}$ to be the homomorphism that sends $g \in G$ to its inverse $g^{-1} \in G^{\mathrm{op}}$; this mapping does not define an automorphism of G, because it fails to commute with the group multiplication, but it does define a homomorphism $G \to G^{\mathrm{op}}$. Now given any homomorphism $\phi \colon G \to H$, the diagram

$$\begin{tikzcd} G \arrow[r, "\eta_G"] \arrow[d, "\phi"'] & G^{\mathrm{op}} \arrow[d, "\phi^{\mathrm{op}}"] \\ H \arrow[r, "\eta_H"'] & H^{\mathrm{op}} \end{tikzcd}$$

commutes because $\phi^{\mathrm{op}}(g^{-1}) = \phi(g^{-1}) = \phi(g)^{-1}$.

(vii) Define an endofunctor of $\mathsf{Vect}_{\Bbbk}$ by $V \mapsto V \otimes V$. There is a natural transformation from the identity functor to this endofunctor whose components are the zero maps, but this is the only such natural transformation: there is no basis-independent way to define a linear map $V \to V \otimes V$. The same result is true for the category of Hilbert

[31] A more flexible notion of *extranatural transformation* can accommodate functors with conflicting variance [**ML98a**, IX.4]; see Exercise 1.4.vi.

[32] A proof that there exists no extranatural isomorphism between the identity and dual functors on the categories of finite-dimensional vector spaces is given in [**EM45**, p. 234].

spaces and linear operators between them, in which context it is related to the "no cloning theorem" in quantum physics.[33]

Another familiar isomorphism that is not natural arises in the classification of finitely generated abelian groups, objects of a category $\mathsf{Ab}_{\mathrm{fg}}$. Let TA denote the **torsion subgroup** of an abelian group A, the subgroup of elements with finite order. In classifying finitely generated groups, one proves that every finitely generated abelian group A is isomorphic to the direct sum $TA \oplus (A/TA)$, the summand A/TA being the **torsion-free** part of A. However, these isomorphisms are not natural, as we now demonstrate.

PROPOSITION 1.4.4. *The isomorphisms $A \cong TA \oplus (A/TA)$ are not natural*[34] *in $A \in \mathsf{Ab}_{\mathrm{fg}}$.*

PROOF. Suppose the isomorphisms $A \cong TA \oplus (A/TA)$ were natural in A. Then the composite

$$A \twoheadrightarrow A/TA \rightarrowtail TA \oplus (A/TA) \cong A \tag{1.4.5}$$

of the canonical quotient map, the inclusion into the direct sum, and the hypothesized natural isomorphism would define a natural endomorphism of the identity functor on $\mathsf{Ab}_{\mathrm{fg}}$. We shall see that this is impossible.

To derive the contradiction, we first show that every natural endomorphism of the identity functor on $\mathsf{Ab}_{\mathrm{fg}}$ is multiplication by some $n \in \mathbb{Z}$. Clearly the component of $\alpha\colon 1_{\mathsf{Ab}_{\mathrm{fg}}} \Rightarrow 1_{\mathsf{Ab}_{\mathrm{fg}}}$ at $\mathbb{Z}$ has this description for some n. Now observe that homomorphisms $\mathbb{Z} \xrightarrow{a} A$ correspond bijectively to elements $a \in A$, choosing a to be the image of $1 \in \mathbb{Z}$. Thus, commutativity of

$$\begin{tikzcd}
\mathbb{Z} \arrow[r, "\alpha_{\mathbb{Z}} = n\cdot -"] \arrow[d, "a"'] & \mathbb{Z} \arrow[d, "a"] \\
A \arrow[r, "\alpha_A"'] & A
\end{tikzcd}$$

forces us to define $\alpha_A(a) = n \cdot a$.

In the case where α is the natural transformation defined by (1.4.5), by examining the component at $A = \mathbb{Z}$, we can see that $n \neq 0$. Finally, consider $A = \mathbb{Z}/2n\mathbb{Z}$. This group is torsion, so any map, such as $\alpha_{\mathbb{Z}/2n\mathbb{Z}}$, which factors through the quotient by its torsion subgroup is zero. But $n \neq 0 \in \mathbb{Z}/2n\mathbb{Z}$, a contradiction. □

EXAMPLE 1.4.6. The Riesz representation theorem can be expressed as a natural isomorphism of functors from the category cHaus of compact Hausdorff spaces and continuous maps to the category Ban of real Banach spaces and continuous linear maps. Let $\Sigma\colon \mathsf{cHaus} \to \mathsf{Ban}$ be the functor that carries a compact Hausdorff space X to the Banach space $\Sigma(X)$ of signed Baire measures on X and sends a continuous map $f\colon X \to Y$ to the map $\mu \mapsto \mu \circ f^{-1}\colon \Sigma(X) \to \Sigma(Y)$. Let $C^*\colon \mathsf{cHaus} \to \mathsf{Ban}$ be the functor that carries X to the linear dual $C(X)^*$ of the Banach space $C(X)$ of continuous real-valued functions on X.

Now for each $\mu \in \Sigma(X)$, there is a linear functional $\phi_\mu\colon C(X) \to \mathbb{R}$ defined by

$$\phi_\mu(g) := \int_X g\, d\mu, \qquad g \in C(X).$$

[33]The states in a quantum mechanical system are modeled by vectors in a Hilbert space and the observables are operators on that space. See [**Bae06**] for more.

[34]Any finitely generated abelian group A has a short exact sequence $0 \to TA \to A \to A/TA \to 0$. Proposition 1.4.4 asserts that there is no natural splitting.

For each $\mu \in \Sigma(X)$, $f : X \to Y$, and $h \in C(Y)$ the equation

$$\int_X hf \, d\mu = \int_Y h \, d(\mu \circ f^{-1})$$

says that the assignment $\mu \mapsto \phi_\mu$ defines the components of a natural transformation $\eta : \Sigma \Rightarrow C^*$. The **Riesz representation theorem** asserts that this natural transformation is a natural isomorphism; see [**Har83**].

EXAMPLE 1.4.7. Consider morphisms $f : w \to x$ and $h : y \to z$ in a locally small category C. Post-composition by h and pre-composition by f define functions between hom-sets

$$\begin{tikzcd} \mathsf{C}(x,y) \arrow[r, "h\cdot -"] \arrow[d, "-\cdot f"'] & \mathsf{C}(x,z) \arrow[d, "-\cdot f"] \\ \mathsf{C}(w,y) \arrow[r, "h\cdot -"'] & \mathsf{C}(w,z) \end{tikzcd} \tag{1.4.8}$$

In Definition 1.3.13 and elsewhere, $h \cdot -$ was denoted by h_* and $- \cdot f$ was denoted by f^*, but we find this less-concise notation to be more evocative here. Associativity of composition implies that this diagram commutes: for any $g : x \to y$, the common image is $hgf : w \to z$.

Interpreting the vertical arrows as the images of f under the actions of the functors $\mathsf{C}(-, y)$ and $\mathsf{C}(-, z)$, the square (1.4.8) demonstrates that there is a natural transformation

$$h_* : \mathsf{C}(-, y) \Rightarrow \mathsf{C}(-, z)$$

whose components are defined by post-composition with $h : y \to z$. Flipping perspectives and interpreting the horizontal arrows as the images of h under the actions of the functors $\mathsf{C}(x, -)$ and $\mathsf{C}(w, -)$, the square (1.4.8) demonstrates that there is a natural transformation

$$f^* : \mathsf{C}(x, -) \Rightarrow \mathsf{C}(w, -)$$

whose components are defined by pre-composition with $f : w \to x$.

A final example describes the natural isomorphisms that supply proofs of the fundamental laws of elementary arithmetic.

EXAMPLE 1.4.9 (a categorification of the natural numbers). For sets A and B, let $A \times B$ denote their cartesian product, let $A + B$ denote their disjoint union, and let A^B denote the set of functions from B to A. These constructions are related by natural isomorphisms

$$\begin{aligned} A \times (B + C) &\cong (A \times B) + (A \times C) & \qquad (A \times B)^C &\cong A^C \times B^C \\ A^{B+C} &\cong A^B \times A^C & (A^B)^C &\cong A^{B \times C} \end{aligned} \tag{1.4.10}$$

In the first instance, the isomorphism defines the components of a natural transformation between a pair of functors $\mathsf{Set} \times \mathsf{Set} \times \mathsf{Set} \to \mathsf{Set}$. For the others, the variance in the variables appearing as "exponents" is contravariant. This is because the assignment $(B, A) \mapsto A^B$ defines a functor $\mathsf{Set}^{\mathrm{op}} \times \mathsf{Set} \to \mathsf{Set}$, namely the two-sided represented functor introduced in Definition 1.3.13.

The displayed natural isomorphisms restrict to the category $\mathsf{Fin}_{\mathrm{iso}}$ of finite sets and bijections, a category which serves as the domain for the cardinality functor $|-| : \mathsf{Fin}_{\mathrm{iso}} \to \mathbb{N}$, whose codomain is the discrete category of natural numbers.[35] Writing $a = |A|$, $b = |B|$,

[35] Mathematical invariants often take the form of a functor from a groupoid to a discrete category.

and $c = |C|$, the cardinality functor carries these natural isomorphisms to the equations

$$a \times (b + c) = (a \times b) + (a \times c) \qquad (a \times b)^c = a^c \times b^c$$
$$a^{b+c} = a^b \times a^c \qquad (a^b)^c = a^{(b\times c)}$$

through a process called **decategorification**. Reversing directions, $\mathsf{Fin}_{\mathrm{iso}}$ is a **categorification** of the natural numbers, which reveals that the familiar laws of arithmetic follow from more fundamental natural isomorphisms between various constructions on sets. A slick proof of each of the displayed natural isomorphisms (1.4.10) appears in Corollary 4.5.6.

Exercises.

EXERCISE 1.4.i. Suppose $\alpha\colon F \Rightarrow G$ is a natural isomorphism. Show that the inverses of the component morphisms define the components of a natural isomorphism $\alpha^{-1}\colon G \Rightarrow F$.

EXERCISE 1.4.ii. What is a natural transformation between a parallel pair of functors between groups, regarded as one-object categories?

EXERCISE 1.4.iii. What is a natural transformation between a parallel pair of functors between preorders, regarded as categories?

EXERCISE 1.4.iv. In the notation of Example 1.4.7, prove that distinct parallel morphisms $f, g\colon c \rightrightarrows d$ define distinct natural transformations

$$f_*, g_*\colon \mathsf{C}(-, c) \Rightarrow \mathsf{C}(-, d) \quad \text{and} \quad f^*, g^*\colon \mathsf{C}(d, -) \Rightarrow \mathsf{C}(c, -)$$

by post- and pre-composition.

EXERCISE 1.4.v. Recall the construction of the comma category for any pair of functors $F\colon \mathsf{D} \to \mathsf{C}$ and $G\colon \mathsf{E} \to \mathsf{C}$ described in Exercise 1.3.vi. From this data, construct a canonical natural transformation $\alpha\colon F\,\mathrm{dom} \Rightarrow G\,\mathrm{cod}$ between the functors that form the boundary of the square

$$\begin{tikzcd} F\downarrow G \arrow[r, "\mathrm{cod}"] \arrow[d, "\mathrm{dom}"'] \arrow[dr, phantom, "\overset{\alpha}{\nearrow}"] & \mathsf{E} \arrow[d, "G"] \\ \mathsf{D} \arrow[r, "F"'] & \mathsf{C} \end{tikzcd}$$

EXERCISE 1.4.vi. Given a pair of functors $F\colon \mathsf{A} \times \mathsf{B} \times \mathsf{B}^{\mathrm{op}} \to \mathsf{D}$ and $G\colon \mathsf{A} \times \mathsf{C} \times \mathsf{C}^{\mathrm{op}} \to \mathsf{D}$, a family of morphisms

$$\alpha_{a,b,c}\colon F(a, b, b) \to G(a, c, c)$$

in D defines the components of an **extranatural transformation** $\alpha\colon F \Rightarrow G$ if for any $f\colon a \to a'$, $g\colon b \to b'$, and $h\colon c \to c'$ the following diagrams commute in D:

$$\begin{tikzcd} F(a,b,b) \arrow[r, "\alpha_{a,b,c}"] \arrow[d, "F(f,1_b,1_b)"'] & G(a,c,c) \arrow[d, "G(f,1_c,1_c)"] \\ F(a',b,b) \arrow[r, "\alpha_{a',b,c}"'] & G(a',c,c) \end{tikzcd} \quad \begin{tikzcd} F(a,b,b') \arrow[r, "F(1_a,1_b,g)"] \arrow[d, "F(1_a,g,1_{b'})"'] & F(a,b,b) \arrow[d, "\alpha_{a,b,c}"] \\ F(a,b',b') \arrow[r, "\alpha_{a,b',c}"'] & G(a,c,c) \end{tikzcd} \quad \begin{tikzcd} F(a,b,b) \arrow[r, "\alpha_{a,b,c'}"] \arrow[d, "\alpha_{a,b,c}"'] & G(a,c',c') \arrow[d, "G(1_a,1_{c'},h)"] \\ G(a,c,c) \arrow[r, "G(1_a,h,1_c)"'] & G(a,c',c) \end{tikzcd}$$

The left-hand square asserts that the components $\alpha_{-,b,c}\colon F(-, b, b) \Rightarrow G(-, c, c)$ define a natural transformation in a for each $b \in \mathsf{B}$ and $c \in \mathsf{C}$. The remaining squares assert that the components $\alpha_{a,-,c}\colon F(a, -, -) \Rightarrow G(a, c, c)$ and $\alpha_{a,b,-}\colon F(a, b, b) \Rightarrow G(a, -, -)$ define transformations that are respectively extranatural in b and in c. Explain why the functors F and G must have a common target category for this definition to make sense.

1.5. Equivalence of categories

> *... la mathématique est l'art de donner le même nom à des choses différentes.*
> ... mathematics is the art of giving the same name to different things.
>
> Henri Poincaré, "L'avenir des mathématiques" [**Poi08**]

There is an analogy between the notion of a natural transformation between a parallel pair of functors and the notion of a homotopy between a parallel pair of continuous functions with one important difference: natural transformations are not generally invertible.[36] As in Example 1.1.4(iv), let $\mathbb{1}$ denote the discrete category with a single object and let $\mathbb{2}$ denote the category with two objects $0, 1 \in \mathbb{2}$ and a single non-identity arrow $0 \to 1$. There are two evident functors $i_0, i_1 \colon \mathbb{1} \rightrightarrows \mathbb{2}$ whose subscripts designate the objects in their image.

LEMMA 1.5.1. *Fixing a parallel pair of functors $F, G \colon \mathsf{C} \rightrightarrows \mathsf{D}$, natural transformations $\alpha \colon F \Rightarrow G$ correspond bijectively to functors $H \colon \mathsf{C} \times \mathbb{2} \to \mathsf{D}$ such that H restricts along i_0 and i_1 to the functors F and G, i.e., so that*

(1.5.2)
$$\begin{tikzcd} \mathsf{C} \arrow[r, "i_0"] \arrow[dr, "F"'] & \mathsf{C} \times \mathbb{2} \arrow[d, "H"] & \mathsf{C} \arrow[l, "i_1"'] \arrow[dl, "G"] \\ & \mathsf{D} & \end{tikzcd}$$

commutes.

Here i_0 denotes the functor defined on objects by $c \mapsto (c, 0)$; it may be regarded as the product of the identity functor on C with the functor i_0.

PROOF. Exercise 1.5.i. $\square$

For instance, if $\mathsf{C} = \mathbb{2}$, each functor $F, G \colon \mathbb{2} \rightrightarrows \mathsf{D}$ picks out an arrow of D, which we also denote by F and G. The directed graph underlying the category $\mathbb{2} \times \mathbb{2}$ is

(1.5.3)
$$\begin{tikzcd} \bullet \arrow[r] \arrow[d] \arrow[dr] & \bullet \arrow[d] \\ \bullet \arrow[r] & \bullet \end{tikzcd}$$

together with four identity endoarrows not depicted here; the diagonal serves as the common composite of the edges of the square. The functor H necessarily maps the top and bottom arrows of (1.5.3) to F and G, respectively. The vertical arrows define the components α_0 and α_1 of the natural transformation, and the diagonal arrow witnesses that the square formed by these four morphisms in D commutes.

If, in (1.5.2), the category $\mathbb{2}$ were replaced by the category $\mathbb{I}$ with two objects and a single arrow in each hom-set, necessarily an isomorphism, then "homotopies" with this interval correspond bijectively to natural *isomorphisms*. The category $\mathbb{2}$ defines the **walking arrow** or **free arrow**, while $\mathbb{I}$ defines the **walking isomorphism** or **free isomorphism**, in a sense that is explained in Examples 2.1.5(x) and (xi).

[36] A natural transformation is invertible if and only if each of its constituent arrows is an isomorphism, in which case the pointwise inverses assemble into a natural transformation by Exercise 1.4.i.

Natural isomorphisms are used to define the notion of *equivalence* of categories. An equivalence of categories is precisely a "homotopy equivalence" where the notion of homotopy is defined using the category $\mathbb{I}$.

DEFINITION 1.5.4. An **equivalence of categories** consists of functors $F\colon \mathsf{C} \leftrightarrows \mathsf{D} \colon G$ together with natural isomorphisms $\eta\colon 1_{\mathsf{C}} \cong GF$, $\epsilon\colon FG \cong 1_{\mathsf{D}}$.[37] Categories C and D are **equivalent**, written $\mathsf{C} \simeq \mathsf{D}$, if there exists an equivalence between them.

Unsurprisingly:

LEMMA 1.5.5. *The notion of equivalence of categories defines an equivalence relation. In particular, if* $\mathsf{C} \simeq \mathsf{D}$ *and* $\mathsf{D} \simeq \mathsf{E}$*, then* $\mathsf{C} \simeq \mathsf{E}$.

PROOF. Exercise 1.5.vi. □

EXAMPLE 1.5.6. The functors $(-)_+\colon \mathsf{Set}^\partial \to \mathsf{Set}_*$ and $U\colon \mathsf{Set}_* \to \mathsf{Set}^\partial$ introduced in §1.3 define an equivalence of categories between the category of pointed sets and the category of sets and partial functions. The composite $U(-)_+$ is the identity on Set^∂, so one of the required natural isomorphisms is the identity. There is a natural isomorphism $\eta\colon 1_{\mathsf{Set}_*} \cong (U-)_+$ whose components

$$\eta_{(X,x)}\colon (X,x) \to (X\backslash\{x\} \cup \{X\backslash\{x\}\}, X\backslash\{x\})$$

are defined to be the based functions that act as the identity on $X\backslash\{x\}$.

Consider the categories $\mathsf{Mat}_{\mathbb{k}}$ and $\mathsf{Vect}^{\mathrm{fd}}_{\mathbb{k}}$ of $\mathbb{k}$-matrices and finite-dimensional non-zero $\mathbb{k}$-vector spaces together with an intermediate category $\mathsf{Vect}^{\mathrm{basis}}_{\mathbb{k}}$ whose objects are finite-dimensional vector spaces with chosen basis and whose morphisms are arbitrary (not necessarily basis-preserving) linear maps. These categories are related by the displayed sequence of functors:

$$\begin{tikzcd} \mathsf{Mat}_{\mathbb{k}} \arrow[r, shift left, "\mathbb{k}^{(-)}"] & \mathsf{Vect}^{\mathrm{basis}}_{\mathbb{k}} \arrow[l, shift left, "H"] \arrow[r, shift left, "U"] & \mathsf{Vect}^{\mathrm{fd}}_{\mathbb{k}} \arrow[l, shift left, "C"] \end{tikzcd}$$

Here $U\colon \mathsf{Vect}^{\mathrm{basis}}_{\mathbb{k}} \to \mathsf{Vect}^{\mathrm{fd}}_{\mathbb{k}}$ is the forgetful functor. The functor $\mathbb{k}^{(-)}\colon \mathsf{Mat}_{\mathbb{k}} \to \mathsf{Vect}^{\mathrm{basis}}_{\mathbb{k}}$ sends n to the vector space $\mathbb{k}^n$, equipped with the standard basis. An $m\times n$-matrix, interpreted with respect to the standard bases on $\mathbb{k}^n$ and $\mathbb{k}^m$, defines a linear map $\mathbb{k}^n \to \mathbb{k}^m$ and this assignment is functorial. The functor H carries a vector space to its dimension and a linear map $\phi\colon V \to W$ to the matrix expressing the action of ϕ on the chosen basis of V using the chosen basis of W. The functor C is defined by choosing a basis for each vector space.

Our aim is to show that these functors display equivalences of categories

$$\mathsf{Mat}_{\mathbb{k}} \simeq \mathsf{Vect}^{\mathrm{basis}}_{\mathbb{k}} \simeq \mathsf{Vect}^{\mathrm{fd}}_{\mathbb{k}}.$$

The composite equivalence $\mathsf{Mat}_{\mathbb{k}} \simeq \mathsf{Vect}^{\mathrm{fd}}_{\mathbb{k}}$ expresses an equivalence between concrete and abstract presentations of linear algebra. A direct proof of these equivalences, by defining suitable natural isomorphisms, is not difficult, but we prefer to give an indirect proof via a useful general theorem characterizing those functors forming part of an equivalence of categories. Its statement requires a few definitions:

DEFINITION 1.5.7. A functor $F\colon \mathsf{C} \to \mathsf{D}$ is

- **full** if for each $x, y \in \mathsf{C}$, the map $\mathsf{C}(x,y) \to \mathsf{D}(Fx,Fy)$ is surjective;
- **faithful** if for each $x, y \in \mathsf{C}$, the map $\mathsf{C}(x,y) \to \mathsf{D}(Fx,Fy)$ is injective;

[37]The notion of equivalence of categories was introduced by Grothendieck in the form of what we would now call an *adjoint equivalence*; this definition appears in Proposition 4.4.5. This explains the directions we have adopted for the natural isomorphisms η and ϵ, which are otherwise immaterial (see Exercise 1.4.i).

- and **essentially surjective on objects** if for every object $d \in \mathsf{D}$ there is some $c \in \mathsf{C}$ such that d is isomorphic to Fc.

Remark 1.5.8. Fullness and faithfulness are *local* conditions; a *global* condition, by contrast, applies "everywhere." A faithful functor need not be injective on morphisms; neither must a full functor be surjective on morphisms. A faithful functor that is injective on objects is called an **embedding** and identifies the domain category as a subcategory of the codomain; in this case, faithfulness implies that the functor is (globally) injective on arrows. A full and faithful functor, called **fully faithful** for short, that is injective-on-objects defines a **full embedding** of the domain category into the codomain category. The domain then defines a **full subcategory** of the codomain.

Theorem 1.5.9 (characterizing equivalences of categories). *A functor defining an equivalence of categories is full, faithful, and essentially surjective on objects. Assuming the axiom of choice, any functor with these properties defines an equivalence of categories.*

The proof of Theorem 1.5.9 makes repeated use of the following elementary lemma.

Lemma 1.5.10. *Any morphism $f: a \to b$ and fixed isomorphisms $a \cong a'$ and $b \cong b'$ determine a unique morphism $f': a' \to b'$ so that any of—or, equivalently, all of—the following four diagrams commute:*

$$\begin{tikzcd} a \arrow[d, "f"'] & a' \arrow[l, "\cong"'] \arrow[d, "f'"] \\ b \arrow[r, "\cong"'] & b' \end{tikzcd} \qquad \begin{tikzcd} a \arrow[r, "\cong"] \arrow[d, "f"'] & a' \arrow[d, "f'"] \\ b \arrow[r, "\cong"'] & b' \end{tikzcd} \qquad \begin{tikzcd} a \arrow[d, "f"'] & a' \arrow[l, "\cong"'] \arrow[d, "f'"] \\ b & b' \arrow[l, "\cong"] \end{tikzcd} \qquad \begin{tikzcd} a \arrow[r, "\cong"] \arrow[d, "f"'] & a' \arrow[d, "f'"] \\ b & b' \arrow[l, "\cong"] \end{tikzcd}$$

Proof. The left-hand diagram defines f'. The commutativity of the remaining diagrams is left as Exercise 1.5.iii. □

Proof[38] of Theorem 1.5.9. First suppose that $F: \mathsf{C} \to \mathsf{D}$, $G: \mathsf{D} \to \mathsf{C}$, $\eta: 1_{\mathsf{C}} \cong GF$, and $\epsilon: FG \cong 1_{\mathsf{D}}$ define an equivalence of categories. For any $d \in \mathsf{D}$, the component of the natural isomorphism $\epsilon_d: FGd \cong d$ demonstrates that F is essentially surjective. Consider a parallel pair $f, g: c \rightrightarrows c'$ in C. If $Ff = Fg$, then both f and g define an arrow $c \to c'$ making the diagram

$$\begin{tikzcd} c \arrow[r, "\eta_c", "\cong"'] \arrow[d, "f \text{ or } g"'] & GFc \arrow[d, "GFf=GFg"] \\ c' \arrow[r, "\cong", "\eta_{c'}"'] & GFc' \end{tikzcd}$$

that expresses the naturality of η commute. Lemma 1.5.10 implies that there is a unique arrow $c \to c'$ with this property, whence $f = g$. Thus, F is faithful, and by symmetry, so is G. Given $k: Fc \to Fc'$, by Lemma 1.5.10, Gk and the isomorphisms η_c and $\eta_{c'}$ define a unique $h: c \to c'$ for which both Gk and GFh make the diagram

$$\begin{tikzcd} c \arrow[r, "\eta_c", "\cong"'] \arrow[d, dashed, "h"'] & GFc \arrow[d, "Gk \text{ or } GFh"] \\ c' \arrow[r, "\cong", "\eta_{c'}"'] & GFc' \end{tikzcd}$$

[38]The reader is strongly encouraged to stop reading here and attempt to prove this result on their own. Indeed, in the first iteration of the course that produced these lecture notes, this proof was left to the homework, without even the time-saving suggestion of Lemma 1.5.10.

commute. By Lemma 1.5.10 again, $GFh = Gk$, whence $Fh = k$ by faithfulness of G. Thus, F is full, faithful, and essentially surjective.

For the converse, suppose now that $F\colon \mathsf{C} \to \mathsf{D}$ is full, faithful, and essentially surjective on objects. Using essential surjectivity and the axiom of choice, choose, for each $d \in \mathsf{D}$, an object $Gd \in \mathsf{C}$ and an isomorphism $\epsilon_d\colon FGd \cong d$. For each $\ell\colon d \to d'$, Lemma 1.5.10 defines a unique morphism making the square

$$\begin{tikzcd} FGd \arrow[r, "\epsilon_d", "\cong"'] \arrow[d, dashed] & d \arrow[d, "\ell"] \\ FGd' \arrow[r, "\cong", "\epsilon_{d'}"'] & d' \end{tikzcd}$$

commute. Since F is fully faithful, there is a unique morphism $Gd \to Gd'$ with this image under F, which we define to be $G\ell$. This definition is arranged so that the chosen isomorphisms assemble into the components of a natural transformation $\epsilon\colon FG \Rightarrow 1_{\mathsf{D}}$. It remains to prove that the assignment of arrows $\ell \mapsto G\ell$ is functorial and to define the natural isomorphism $\eta\colon 1_{\mathsf{C}} \Rightarrow GF$.

Functoriality of G is another consequence of Lemma 1.5.10 and faithfulness of F. The morphisms $FG1_d$ and $F1_{Gd}$ both make

$$\begin{tikzcd} FGd \arrow[r, "\epsilon_d", "\cong"'] \arrow[d, "FG1_d \text{ or } F1_{Gd}"'] & d \arrow[d, "1_d"] \\ FGd \arrow[r, "\cong", "\epsilon_d"'] & d \end{tikzcd}$$

commute, whence $G1_d = 1_{Gd}$. Similarly, given $\ell'\colon d' \to d''$, both $F(G\ell' \cdot G\ell)$ and $FG(\ell'\ell)$ make

$$\begin{tikzcd} FGd \arrow[r, "\epsilon_d", "\cong"'] \arrow[d, "F(G\ell' \cdot G\ell) \text{ or } FG(\ell'\ell)"'] & d \arrow[d, "\ell'\ell"] \\ FGd'' \arrow[r, "\cong", "\epsilon_{d''}"'] & d'' \end{tikzcd}$$

commute, whence $G\ell' \cdot G\ell = G(\ell'\ell)$.

Finally, by full and faithfulness of F, we may define the isomorphisms $\eta_c\colon c \to GFc$ by specifying isomorphisms $F\eta_c\colon Fc \to FGFc$; see Exercise 1.5.iv. Define $F\eta_c$ to be ϵ_{Fc}^{-1}. For any $f\colon c \to c'$, the outer rectangle

$$\begin{tikzcd} Fc \arrow[r, "F\eta_c"] \arrow[d, "Ff"'] & FGFc \arrow[r, "\epsilon_{Fc}"] \arrow[d, "FGFf"'] & Fc \arrow[d, "Ff"] \\ Fc' \arrow[r, "F\eta_{c'}"'] & FGFc' \arrow[r, "\epsilon_{Fc'}"'] & Fc' \end{tikzcd}$$

commutes, both composites being Ff. The right-hand square commutes by naturality of ϵ. Because $\epsilon_{Fc'}$ is an isomorphism, this implies that the left-hand square commutes; see Lemma 1.6.21. Faithfulness of F tells us that $\eta_{c'} \cdot f = GFf \cdot \eta_c$, i.e., that η is a natural transformation. □

The proof of Theorem 1.5.9 is an example of a proof by "diagram chasing," a technique that is introduced more formally in §1.6. If some of the steps were hard to follow, we suggest having a second look after that section is absorbed.

COROLLARY 1.5.11 (an equivalence between abstract and concrete linear algebra). *For any field $\Bbbk$, the categories $\mathsf{Mat}_{\Bbbk}$ and $\mathsf{Vect}^{\mathrm{fd}}_{\Bbbk}$ are equivalent.*

PROOF. Applying Theorem 1.5.9, it is easy to see that each of the functors

$$\mathsf{Mat}_{\Bbbk} \leftrightarrows \mathsf{Vect}^{\mathrm{basis}}_{\Bbbk} \rightleftarrows \mathsf{Vect}^{\mathrm{fd}}_{\Bbbk}$$

defines an equivalence of categories. For instance, the morphisms in the category $\mathsf{Vect}^{\mathrm{basis}}_{\Bbbk}$ are defined so that $U \colon \mathsf{Vect}^{\mathrm{basis}}_{\Bbbk} \to \mathsf{Vect}^{\mathrm{fd}}_{\Bbbk}$ is fully faithful. □

A category is **connected** if any pair of objects can be connected by a finite zig-zag of morphisms.

PROPOSITION 1.5.12. *Any connected groupoid is equivalent, as a category, to the automorphism group of any of its objects.*

PROOF. Choose any object g of a connected groupoid G and let $G = \mathsf{G}(g, g)$ denote its automorphism group. The inclusion $\mathsf{B}G \hookrightarrow \mathsf{G}$ mapping the unique object of $\mathsf{B}G$ to $g \in \mathsf{G}$ is full and faithful, by definition, and essentially surjective, since G was assumed to be connected. Apply Theorem 1.5.9. □

As a special case, we obtain the following result:

COROLLARY 1.5.13. *In a path-connected space X, any choice of basepoint $x \in X$ yields an isomorphic fundamental group $\pi_1(X, x)$.*

PROOF. Recall from Example 1.1.12(ii) that any space X has a fundamental groupoid $\Pi_1(X)$ whose objects are points in X and whose morphisms are endpoint-preserving homotopy classes of paths in X. Picking any point x, the group of automorphisms of the object $x \in \Pi_1(X)$ is exactly the fundamental group $\pi_1(X, x)$. Proposition 1.5.12 implies that any pair of automorphism groups are equivalent, as categories, to the fundamental groupoid

$$\pi_1(X,x) \xhookrightarrow{\sim} \Pi_1(X) \xhookleftarrow{\sim} \pi_1(X,x')$$

and thus to each other. An equivalence between 1-object categories is an isomorphism. Exercise 1.3.i reveals that an isomorphism of groups, regarded as 1-object categories, is exactly an isomorphism of groups in the usual sense (a bijective homomorphism). Thus, all of the fundamental groups defined by choosing a basepoint in a path-connected space are isomorphic. □

REMARK 1.5.14. Frequently, one functor of an equivalence of categories can be defined canonically, while the inverse equivalence requires the axiom of choice. In the case of the equivalence between the fundamental group and fundamental groupoid of a path-connected space, a more precise statement is that one comparison equivalence is natural while the other, making use of the axiom of choice, is not. Write $\mathsf{Top}^{\mathrm{pc}}_{*}$ for the category of path-connected based topological spaces. We regard the fundamental group π_1 and fundamental groupoid Π_1 as a parallel pair of functors:

$$\pi_1 \colon \mathsf{Top}^{\mathrm{pc}}_{*} \xrightarrow{\pi_1} \mathsf{Group} \hookrightarrow \mathsf{Cat} \qquad \text{and} \qquad \Pi_1 \colon \mathsf{Top}^{\mathrm{pc}}_{*} \xrightarrow{U} \mathsf{Top} \xrightarrow{\Pi_1} \mathsf{Groupoid} \hookrightarrow \mathsf{Cat}\,.$$

The inclusion of the fundamental group into the fundamental groupoid defines a natural transformation $\pi_1 \Rightarrow \Pi_1$ such that each component $\pi_1(X, x) \to \Pi_1(X)$, itself a functor, is furthermore an equivalence of categories. The definition of the inverse equivalence $\Pi_1(X) \to \pi_1(X, x)$ requires the choice, for each point $p \in X$, of a path-connecting p to the basepoint x. These (path homotopy classes of) chosen paths need not be preserved by maps in $\mathsf{Top}_*^{\mathrm{pc}}$. Thus, the inverse equivalences $\Pi_1(X) \to \pi_1(X, x)$ do not assemble into a natural transformation.

The group of automorphisms of any object in a connected groupoid, considered in Proposition 1.5.12, is one example of a *skeleton* of a category.

Definition 1.5.15. A category C is **skeletal** if it contains just one object in each isomorphism class. The **skeleton** skC of a category C is the unique (up to isomorphism) skeletal category that is equivalent to C.

Remark 1.5.16. The category skC may be constructed by choosing one object in each isomorphism class in C and defining skC to be the full subcategory on this collection of objects. By Theorem 1.5.9, the inclusion $\mathsf{skC} \hookrightarrow \mathsf{C}$ defines an equivalence of categories. This construction, however, fails to define a functor $\mathsf{sk}(-)\colon \mathsf{CAT} \to \mathsf{CAT}$ because there is no reason that a functor $F\colon \mathsf{C} \to \mathsf{D}$ would necessarily restrict to define a functor between the chosen skeletal subcategories. It is possible to choose a functor $\mathsf{sk}F\colon \mathsf{skC} \to \mathsf{skD}$ whose inclusion into D is naturally isomorphic to the restriction of F to skC, but these choices will not be strictly functorial.[39]

Note than an equivalence between skeletal categories is necessarily an isomorphism of categories. Thus, two categories are equivalent if and only if their skeletons are isomorphic. For this reason, we feel free to speak of *the* skeleton of a category, even though its construction is not canonical.

Example 1.5.17.

(i) The skeleton of a connected groupoid is the group of automorphisms of any of its objects (see Proposition 1.5.12).
(ii) The skeleton of the category defined by a preorder, as described in Example 1.1.4(iii), is a poset.
(iii) The skeleton of the category $\mathsf{Vect}_{\Bbbk}^{\mathrm{fd}}$ is the category $\mathsf{Mat}_{\Bbbk}$.
(iv) The skeleton of the category $\mathsf{Fin}_{\mathrm{iso}}$ is the category whose objects are positive integers and with $\mathrm{Hom}(n, n) = \Sigma_n$, the group of permutations of n elements. The hom-sets between distinct natural numbers are all empty.

Example 1.5.18 (a categorification of the orbit-stabilizer theorem). Let $X\colon \mathsf{B}G \to \mathsf{Set}$ be a left G-set. Its **translation groupoid** $\mathsf{T}_G X$ has elements of X as objects. A morphism $g\colon x \to y$ is an element $g \in G$ so that $g \cdot x = y$. The objects in the skeleton $\mathsf{skT}_G X$ are the connected components in the translation groupoid. These are precisely the **orbits** of the group action, which partition X in precisely this manner.

Consider $x \in X$ as a representative of its orbit O_x. Because the translation groupoid is equivalent to its skeleton, we must have

$$\mathrm{Hom}_{\mathsf{skT}_G X}(O_x, O_x) \cong \mathrm{Hom}_{\mathsf{T}_G X}(x, x) =: G_x\,,$$

the set of automorphisms of x. This group consists of precisely those $g \in G$ so that $g \cdot x = x$. In other words, the group $\mathrm{Hom}_{\mathsf{T}_G X}(x, x)$ is the **stabilizer** G_x of x with respect to the G-action. Note that this argument implies that any pair of elements in the same orbit must have isomorphic stabilizers. As is always the case for a skeletal groupoid, there are no morphisms between distinct objects. In summary, the skeleton of the translation groupoid, as a category, is the disjoint union of the stabilizer groups, indexed by the orbits of the action of G on X.

---

[39]There is, however, a *pseudofunctor* $\mathsf{sk}(-)\colon \mathsf{CAT} \to \mathsf{CAT}$, but a precise definition of this concept is beyond the scope of this text.

The set of morphisms in the translation groupoid with domain x is isomorphic to G. This set may be expressed as a disjoint union of hom-sets $\mathrm{Hom}_{\mathsf{T}_G X}(x, y)$, where y ranges over the orbit O_x. Each of these hom-sets is isomorphic to $\mathrm{Hom}_{\mathsf{T}_G X}(x, x) = G_x$. In particular, $|G| = |O_x| \cdot |G_x|$, proving the **orbit-stabilizer theorem**.

A guiding principle in category theory is that categorically-defined concepts should be equivalence invariant. Some category theorists go so far as to call a definition "evil" if it is not invariant under equivalence of categories. The only evil definitions that have been introduced thus far are smallness and discreteness. A category is **essentially small** if it is equivalent to a small category or, equivalently, if its skeleton is a small category. A category is **essentially discrete** if it is equivalent to a discrete category.

The following constructions and definitions are equivalence invariant:

- If a category is locally small, any category equivalent to it is again locally small.
- If a category is a groupoid, any category equivalent to it is again a groupoid.
- If $\mathsf{C} \simeq \mathsf{D}$, then $\mathsf{C}^{\mathrm{op}} \simeq \mathsf{D}^{\mathrm{op}}$.
- The product of a pair of categories is equivalent to the product of any pair of equivalent categories.
- An arrow in C is an isomorphism if and only if its image under an equivalence $\mathsf{C} \xrightarrow{\simeq} \mathsf{D}$ is an isomorphism.

The last of these properties can be generalized. By Theorem 1.5.9, a full and faithful functor $F\colon \mathsf{C} \to \mathsf{D}$ defines an equivalence onto its **essential image**, the full subcategory of objects isomorphic to Fc for some $c \in \mathsf{C}$. Fully faithful functors have a useful property stated as Exercise 1.5.iv: if F is full and faithful and Fc and Fc' are isomorphic in D, then c and c' are isomorphic in C. We will introduce what are easily the most important fully faithful functors in category theory in Chapter 2: the covariant and contravariant Yoneda embeddings.

Exercises.

EXERCISE 1.5.i. Prove Lemma 1.5.1.

EXERCISE 1.5.ii. Segal defined a category Γ in [**Seg74**] as follows:

> Γ is the category whose objects are all finite sets, and whose morphisms from S to T are the maps $\theta\colon S \to P(T)$ such that $\theta(\alpha)$ and $\theta(\beta)$ are disjoint when $\alpha \neq \beta$. The composite of $\theta\colon S \to P(T)$ and $\phi\colon T \to P(U)$ is $\psi\colon S \to P(U)$, where $\psi(\alpha) = \bigcup_{\beta \in \theta(\alpha)} \phi(\beta)$.

Prove that Γ is equivalent to the opposite of the category Fin_* of finite pointed sets. In particular, the functors introduced in Example 1.3.2(xi) define presheaves on Γ.

EXERCISE 1.5.iii. Prove Lemma 1.5.10.

EXERCISE 1.5.iv. Show that a full and faithful functor $F\colon \mathsf{C} \to \mathsf{D}$ both **reflects** and **creates isomorphisms**. That is, show:

(i) If f is a morphism in C so that Ff is an isomorphism in D, then f is an isomorphism.
(ii) If x and y are objects in C so that Fx and Fy are isomorphic in D, then x and y are isomorphic in C.

By Lemma 1.3.8, the converses of these statements hold for any functor.

EXERCISE 1.5.v. Find an example to show that a faithful functor need not reflect isomorphisms.

EXERCISE 1.5.vi.

(i) Prove that the composite of a pair of full, faithful, or essentially surjective functors again has the same properties.
(ii) Prove that if $\mathsf{C} \simeq \mathsf{D}$ and $\mathsf{D} \simeq \mathsf{E}$, then $\mathsf{C} \simeq \mathsf{E}$. Conclude that equivalence of categories is an equivalence relation.[40]

EXERCISE 1.5.vii. Let G be a connected groupoid and let G be the group of automorphisms at any of its objects. The inclusion $\mathsf{B}G \hookrightarrow \mathsf{G}$ defines an equivalence of categories. Construct an inverse equivalence $\mathsf{G} \to \mathsf{B}G$.

EXERCISE 1.5.viii. Klein's Erlangen program studies groupoids of geometric spaces of various kinds. Prove that the groupoid Affine of affine planes is equivalent to the groupoid $\mathsf{Proj}^|$ of projective planes with a distinguished line, called the "line at infinity." The morphisms in each groupoid are bijections on both points and lines (preserving the distinguished line in the case of projective planes) that preserve and reflect the incidence relation. The functor $\mathsf{Proj}^| \to \mathsf{Affine}$ removes the line at infinity and the points it contains. Explicitly describe an inverse equivalence.

EXERCISE 1.5.ix. Show that any category that is equivalent to a locally small category is locally small.

EXERCISE 1.5.x. Characterize the categories that are equivalent to discrete categories. A category that is connected and essentially discrete is called **chaotic**.

EXERCISE 1.5.xi. Consider the functors $\mathsf{Ab} \to \mathsf{Group}$ (inclusion), $\mathsf{Ring} \to \mathsf{Ab}$ (forgetting the multiplication), $(-)^\times : \mathsf{Ring} \to \mathsf{Group}$ (taking the group of units), $\mathsf{Ring} \to \mathsf{Rng}$ (inclusion), $\mathsf{Field} \to \mathsf{Ring}$ (inclusion), and $\mathsf{Mod}_R \to \mathsf{Ab}$ (forgetful). Determine which functors are full, which are faithful, and which are essentially surjective. Do any define an equivalence of categories? (Warning: A few of these questions conceal research-level problems, but they can be fun to think about even if full solutions are hard to come by.)

1.6. The art of the diagram chase

> The diagrams incorporate a large amount of information. Their use provides extensive savings in space and in mental effort. In the case of many theorems, the setting up of the correct diagram is the major part of the proof. We therefore urge that the reader stop at the end of each theorem and attempt to construct for himself the relevant diagram before examining the one which is given in the text. Once this is done, the subsequent demonstration can be followed more readily; in fact, the reader can usually supply it himself.
>
> Samuel Eilenberg and Norman Steenrod, *Foundations of Algebraic Topology* [**ES52**]

The proof of Theorem 1.5.9 used a technique called "diagram chasing," also called *abstract nonsense*, that we now consider more formally. A **diagram** is typically presented informally as a directed graph of morphisms in a category. In this informal presentation, the diagram **commutes** if any two paths of composable arrows in the directed graph with

[40]A second, more direct proof of this result appears as Exercise 1.7.vi.

common source and target have the same composite. For example, a commutative triangle

(1.6.1)
$$\begin{tikzcd} \bullet \arrow[rr, "h"] \arrow[dr, "f"'] & & \bullet \\ & \bullet \arrow[ur, "g"'] & \end{tikzcd}$$

asserts that the hypotenuse h equals the composite gf of the two legs. Commutative diagrams in a category can be used to define more complicated mathematical objects. For example:

Definition 1.6.2. A **monoid** is an object $M \in \mathsf{Set}$ together with a pair of morphisms $\mu\colon M \times M \to M$ and $\eta\colon 1 \to M$ so that the following diagrams commute:

$$\begin{tikzcd} M \times M \times M \arrow[r, "1_M \times \mu"] \arrow[d, "\mu \times 1_M"'] & M \times M \arrow[d, "\mu"] \\ M \times M \arrow[r, "\mu"'] & M \end{tikzcd} \qquad \begin{tikzcd} M \arrow[r, "\eta \times 1_M"] \arrow[dr, "1_M"'] & M \times M \arrow[d, "\mu"] & M \arrow[l, "1_M \times \eta"'] \arrow[dl, "1_M"] \\ & M & \end{tikzcd}$$

The morphism $\mu\colon M \times M \to M$ defines a binary "multiplication" operation on M. The morphism $\eta\colon 1 \to M$, whose domain is a singleton set, identifies an element $\eta \in M$. The three axioms demand that multiplication is associative and that multiplication on the left or right by the element η acts as the identity. The advantage of the commutative diagrams approach to this definition is that it readily generalizes to other categories. For example:

Definition 1.6.3. A **topological monoid** is an object $M \in \mathsf{Top}$ together with morphisms $\mu\colon M \times M \to M$ and $\eta\colon 1 \to M$ so that the following diagrams commute:

$$\begin{tikzcd} M \times M \times M \arrow[r, "1_M \times \mu"] \arrow[d, "\mu \times 1_M"'] & M \times M \arrow[d, "\mu"] \\ M \times M \arrow[r, "\mu"'] & M \end{tikzcd} \qquad \begin{tikzcd} M \arrow[r, "\eta \times 1_M"] \arrow[dr, "1_M"'] & M \times M \arrow[d, "\mu"] & M \arrow[l, "1_M \times \eta"'] \arrow[dl, "1_M"] \\ & M & \end{tikzcd}$$

A **unital ring**[41] is an object $R \in \mathsf{Ab}$ together with morphisms $\mu\colon R \otimes_{\mathbb{Z}} R \to R$ and $\eta\colon \mathbb{Z} \to R$ so that the following diagrams commute:

$$\begin{tikzcd} R \otimes_{\mathbb{Z}} R \otimes_{\mathbb{Z}} R \arrow[r, "1_R \otimes_{\mathbb{Z}} \mu"] \arrow[d, "\mu \otimes_{\mathbb{Z}} 1_R"'] & R \otimes_{\mathbb{Z}} R \arrow[d, "\mu"] \\ R \otimes_{\mathbb{Z}} R \arrow[r, "\mu"'] & R \end{tikzcd} \qquad \begin{tikzcd} R \arrow[r, "\eta \otimes_{\mathbb{Z}} 1_R"] \arrow[dr, "1_R"'] & R \otimes_{\mathbb{Z}} R \arrow[d, "\mu"] & R \arrow[l, "1_R \otimes_{\mathbb{Z}} \eta"'] \arrow[dl, "1_R"] \\ & R & \end{tikzcd}$$

A $\Bbbk$-**algebra** is an object $R \in \mathsf{Vect}_{\Bbbk}$ together with morphisms $\mu\colon R \otimes_{\Bbbk} R \to R$ and $\eta\colon \Bbbk \to R$ so that the following diagrams commute:

$$\begin{tikzcd} R \otimes_{\Bbbk} R \otimes_{\Bbbk} R \arrow[r, "1_R \otimes_{\Bbbk} \mu"] \arrow[d, "\mu \otimes_{\Bbbk} 1_R"'] & R \otimes_{\Bbbk} R \arrow[d, "\mu"] \\ R \otimes_{\Bbbk} R \arrow[r, "\mu"'] & R \end{tikzcd} \qquad \begin{tikzcd} R \arrow[r, "\eta \otimes_{\Bbbk} 1_R"] \arrow[dr, "1_R"'] & R \otimes_{\Bbbk} R \arrow[d, "\mu"] & R \arrow[l, "1_R \otimes_{\Bbbk} \eta"'] \arrow[dl, "1_R"] \\ & R & \end{tikzcd}$$

There are evident formal similarities in each of these four definitions; they are all special cases of a general notion of a monoid in a *monoidal category*, as defined in §E.2.

[41] A not-necessarily unital ring may be defined by ignoring the morphism η and the pair of commutative triangles.

The morphisms $\eta\colon 1 \to M$ in the case of topological monoids, $\eta\colon \mathbb{Z} \to R$ in the case of unital rings, and $\eta\colon \Bbbk \to R$ in the case of $\Bbbk$-algebras do no more and no less than specify an element of M or R to serve as the multiplicative unit. We will introduce language to describe the role played in each case by the topological space 1, the abelian group $\mathbb{Z}$, and the vector space $\Bbbk$ in Example 2.1.5(iii).

In the case of a topological monoid, the condition that $\mu\colon M\times M \to M$ is a morphism in Top demands that the multiplication function is continuous; for instance, the circle $S^1 \subset \mathbb{C}$ defines a topological monoid with addition of angles. For unital rings, the morphism $\mu\colon R \otimes_{\mathbb{Z}} R \to R$ represents a bilinear homomorphism of abelian groups from $R \times R$ to R; in particular, multiplication distributes over addition in R. The role of the tensor product in the definition of a $\Bbbk$-algebra is similar.

Let us now give precise meaning to the term "diagram."

Definition 1.6.4. A **diagram** in a category C is a functor $F\colon \mathsf{J} \to \mathsf{C}$ whose domain, the **indexing category**, is a small category.

A diagram is typically depicted by drawing the objects and morphisms in its image, with the domain category left implicit, particularly in the case where the indexing category J is a preorder, so that any two paths of composable arrows have a common composite. Nonetheless, the indexing category J plays an important role. Functoriality requires that any composition relations that hold in J must hold in the image of the diagram, which is what it means to say that the directed graph defined by the image of the diagram in C is **commutative**. An immediate consequence of the form of our definition of a commutative diagram is the following result.

Lemma 1.6.5. *Functors preserve commutative diagrams.*

Proof. A diagram in C is given by a functor $F\colon \mathsf{J} \to \mathsf{C}$, whose domain is a small category. Given any functor $G\colon \mathsf{C} \to \mathsf{D}$, the composite $GF\colon \mathsf{J} \to \mathsf{D}$ defines the image of the diagram in D. □

A few examples will help to illustrate the connection between commutative diagrams, as formalized in Definition 1.6.4, and their informal directed graph presentations.

Example 1.6.6. Consider $\mathbb{2} \times \mathbb{2}$, the category with four objects and the displayed non-identity morphisms

$$\begin{tikzcd} \bullet \arrow[r] \arrow[d] \arrow[dr] & \bullet \arrow[d] \\ \bullet \arrow[r] & \bullet \end{tikzcd}$$

In $\mathbb{2}\times\mathbb{2}$, the diagonal morphism is the composite of both the top and right morphisms and the left and bottom morphisms; in particular, these composites are equal. A diagram indexed by $\mathbb{2} \times \mathbb{2}$, typically drawn without the diagonal composite, is a **commutative square**.

Remark 1.6.7. In practice, one thinks of the indexing category as a directed graph, defining the **shape** of the diagram, together with specified commutativity relations. For example, to define a functor with domain $\mathbb{2} \times \mathbb{2}$, it suffices to specify the images of the four objects together with four morphisms

$$\begin{tikzcd} a \arrow[r, "f"] \arrow[d, "g"'] & b \arrow[d, "h"] \\ c \arrow[r, "k"'] & d \end{tikzcd}$$

subject to the relation that $hf = kg$. When indexing categories are represented in this way, the commutativity relations become an essential part of the data. They distinguish between the category $\mathbb{2} \times \mathbb{2}$ that indexes a commutative square and the category

$$\begin{tikzcd} \bullet \arrow[r] \arrow[d] \arrow[dr, shift left] \arrow[dr, shift right] & \bullet \arrow[d] \\ \bullet \arrow[r] & \bullet \end{tikzcd}$$

that indexes a not-necessarily commutative square; here the two diagonals represent distinct composites of the two paths along the edges of the square.

EXAMPLE 1.6.8. The category $\mathbb{4}$ has four objects and the six displayed non-identity morphisms

$$\begin{tikzcd} & \bullet \arrow[dr] \arrow[dd] & \\ \bullet \arrow[ur] \arrow[rr] \arrow[dr] & & \bullet \\ & \bullet \arrow[ur] & \end{tikzcd}$$

This category is a preorder; each of the four displayed triangular "faces" commutes. These commutativity relations in $\mathbb{4}$ imply that a diagram of shape $\mathbb{4}$ in a category C is given by a sequence of three composable morphisms together with their composites

$$\begin{tikzcd} & \bullet \arrow[dr, "hg"] \arrow[dd, "g"'] & \\ \bullet \arrow[ur, "f"] \arrow[rr, "hgf"] \arrow[dr, "gf"'] & & \bullet \\ & \bullet \arrow[ur, "h"'] & \end{tikzcd}$$

Associativity of composition implies that the bottom and back faces commute in C. Example 4.1.13 introduces terminology to explain why a diagram of shape $\mathbb{4}$ is determined by a sequence of three composable morphisms: the category $\mathbb{4}$ is *free on the directed graph* $\bullet \to \bullet \to \bullet \to \bullet$.

EXAMPLE 1.6.9. Consider $\mathbb{2}\times\mathbb{3}$, the category with six objects and the displayed non-identity morphisms

$$\begin{tikzcd} \bullet \arrow[r] \arrow[rr, bend left] \arrow[d] \arrow[dr] \arrow[drr] & \bullet \arrow[r] \arrow[d] \arrow[dr] & \bullet \arrow[d] \\ \bullet \arrow[r] \arrow[rr, bend right] & \bullet \arrow[r] & \bullet \end{tikzcd}$$

The long diagonal asserts that the outer triples of composable morphisms have a common composite, i.e., that the outer rectangle commutes. The short diagonals assert, respectively, that the left-hand and right-hand squares commute. The inner parallelogram also commutes. A diagram indexed by $\mathbb{2}\times\mathbb{3}$, typically drawn without any of the diagonals, is a **commutative rectangle**.

Two commutative squares define a commutative rectangle: a collection of morphisms with the indicated sources and targets

$$\begin{tikzcd} a \arrow[r, "f"] \arrow[d, "g"'] & b \arrow[r, "j"] \arrow[d, "h"] & c \arrow[d, "\ell"] \\ a' \arrow[r, "k"'] & b' \arrow[r, "m"'] & c' \end{tikzcd} \tag{1.6.10}$$

define a $\mathbb{2} \times \mathbb{3}$-shaped diagram provided that $hf = kg$ and $\ell j = mh$. This is a special case of the following more general result, which describes the induced relations in the "algebra of composition" encoded by the arrows in a category.

LEMMA 1.6.11. *Suppose $f_1, \ldots, f_n$ is a composable sequence—a "path"—of morphisms in a category. If the composite $f_k f_{k-1} \cdots f_{i+1} f_i$ equals $g_m \cdots g_1$, for another composable sequence of morphisms $g_1, \ldots, g_m$, then $f_n \cdots f_1 = f_n \cdots f_{k+1} g_m \cdots g_1 f_{i-1} \cdots f_1$.*

PROOF. Composition is well-defined: if the composites $g_m \cdots g_1$ and $f_k f_{k-1} \cdots f_{i+1} f_i$ define the same arrow, then the results of pre- or post-composing with other sequences of arrows must also be the same. □

This very simple result underlies most proofs by "diagram chasing." When a diagram is depicted by a simple (meaning there is at most one edge between any two vertices) acyclic directed graph, the most common convention is to include commutativity relations that assert that any two paths in the diagram with a common source and target commute, i.e., that the directed graph represents a poset category. For example, the category $\mathbb{2} \times \mathbb{2} \times \mathbb{2}$ indexes the **commutative cube**, which is typically depicted as follows:

$$\begin{tikzcd} \bullet \arrow[rr] \arrow[dr] \arrow[dd] & & \bullet \arrow[dr] \arrow[dd] & \\ & \bullet \arrow[rr] \arrow[dd] & & \bullet \arrow[dd] \\ \bullet \arrow[rr, crossing over] \arrow[dr] & & \bullet \arrow[dr] & \\ & \bullet \arrow[rr] & & \bullet \end{tikzcd}$$

In such cases, Lemma 1.6.11 and transitivity of equality implies that commutativity of the entire diagram may be checked by establishing commutativity of each minimal subdiagram in the directed graph. Here, a minimal subdiagram corresponds to a composition relation $h_n \cdots h_1 = k_m \cdots k_1$ that cannot be factored into a relation between shorter paths of composable morphisms. The graph corresponding to a minimal relation is a "directed polygon"

$$\begin{tikzcd}[column sep=small] & \bullet \arrow[r, "h_2"] & \bullet \cdots\cdots \bullet \arrow[r, "h_{n-1}"] & \bullet \arrow[dr, "h_n"] & \\ \bullet \arrow[ur, "h_1"] \arrow[dr, "k_1"'] & & & & \bullet \\ & \bullet \arrow[r, "k_2"'] & \bullet \cdots\cdots \bullet \arrow[r, "k_{m-1}"'] & \bullet \arrow[ur, "k_m"'] & \end{tikzcd}$$

with a commutative triangle, as in (1.6.1), being the simplest case. This sort of argument is called "equational reasoning" in [**Sim11a**, 2.1], which provides an excellent short introduction to diagram chasing.

The following results have simple proofs by diagram chasing.

LEMMA 1.6.12. *If the triangle displayed on the left commutes*

$$\begin{tikzcd} \bullet \arrow[dr, "h"] \arrow[dd, "f"'] & \\ & \bullet \\ \bullet \arrow[ur, "g"'] & \end{tikzcd} \qquad \begin{tikzcd} \bullet \arrow[dr, "h"] & \\ & \bullet \\ \bullet \arrow[ur, "g"'] \arrow[uu, "f^{-1}"] & \end{tikzcd}$$

and if f is an isomorphism, then the triangle displayed on the right also commutes. Dually, for any triple of morphisms with domains and codomains as displayed with k an isomorphism

$$\begin{tikzcd} & \bullet \arrow[dd, "k"] \\ \bullet \arrow[ur, "j"] \arrow[dr, "\ell"'] & \\ & \bullet \end{tikzcd} \qquad \begin{tikzcd} & \bullet \\ \bullet \arrow[ur, "j"] \arrow[dr, "\ell"'] & \\ & \bullet \arrow[uu, "k^{-1}"'] \end{tikzcd}$$

the left-hand triangle commutes if and only if the right-hand triangle commutes.

PROOF. Pre-compose the composition relation $h = gf$ with f^{-1} to yield $hf^{-1} = g$. □

LEMMA 1.6.13. *For any commutative square $\beta\alpha = \delta\gamma$ in which each of the morphisms is an isomorphism, then the inverses define a commutative square $\alpha^{-1}\beta^{-1} = \gamma^{-1}\delta^{-1}$.*

PROOF. Apply Lemma 1.6.12 four times. □

In certain special cases, commutativity of diagrams can be automatic. For instance, any parallel sequences of composable morphisms in a preorder must have a common composite precisely because any hom-set in a preorder has at most one element. Parallel sequences of composable morphisms also necessarily have a common composite when the domain or codomain objects have a certain special property:

DEFINITION 1.6.14. An object $i \in \mathsf{C}$ is **initial** if for every $c \in \mathsf{C}$ there is a unique morphism $i \to c$. Dually, an object $t \in \mathsf{C}$ is **terminal** if for every $c \in \mathsf{C}$ there is a unique morphism $c \to t$.

EXAMPLE 1.6.15. Many of the categories of our acquaintance have initial and terminal objects.

(i) The empty set is an initial object in Set and any singleton set is terminal.
(ii) In Top, the empty and singleton spaces are, respectively, initial and terminal.
(iii) In Set_*, any singleton set is both initial and terminal.
(iv) In Mod_R, the zero module is both initial and terminal. Similarly, the trivial group is both initial and terminal in Group.
(v) The zero ring is the terminal ring. To identify an initial object, we must clarify what sort of rings and ring homomorphisms are being considered. Recall that Ring denotes the category of unital rings and ring homomorphisms that preserve the multiplicative identity. This is a non-full subcategory of the larger category Rng of rings that do not necessarily have a multiplicative identity and homomorphisms that need not preserve one if it happens to exist. The integers define an initial object in Ring but not in Rng, in which the zero ring is initial.
(vi) The category Field of fields has neither initial nor terminal objects. Indeed, there are no homomorphisms between fields of different characteristic.
(vii) The empty category defines an initial object in Cat, and the category $\mathbb{1}$ is terminal.

(viii) An initial object in a preorder is a global minimal element, and a terminal object is a global maximal element. These initial and terminal objects may or may not exist in each particular instance.

LEMMA 1.6.16. *Let $f_1, \ldots, f_n$ and $g_1, \ldots, g_m$ be composable sequences of morphisms so that the domain of f_1 equals the domain of g_1 and the codomain of f_n equals the codomain of g_m. If this common codomain is a terminal object, or if this common domain is an initial object, then $f_n \cdots f_1 = g_m \cdots g_1$.*

PROOF. The two dual statements are immediate consequences of the uniqueness part of Definition 1.6.14. □

In certain cases, one can prove that a diagram commutes by appealing to "elements" of the objects. For instance, this is possible in any concrete category.

DEFINITION 1.6.17. A **concrete category** is a category C equipped with a faithful functor $U\colon \mathsf{C} \to \mathsf{Set}$.

The functor U typically carries an object of C to its "underlying set." The faithfulness condition asserts that any parallel pair of morphisms $f, g\colon c \rightrightarrows c'$ that induce the same function $Uf = Ug$ between the underlying sets must be equal in C. The idea is that the question of whether a map between the underlying sets of objects in a concrete category is a map in the category is a condition (e.g., continuity). By contrast, the functor $U\colon \mathsf{C} \to \mathsf{Set}$ is not faithful if the maps in C have extra structure that is not visible at the level of the underlying sets (e.g., homotopy classes of maps [**Fre04**]).

EXAMPLE 1.6.18. Each category listed in Example 1.1.3 is concrete, although care must be taken in the cases of Graph and Ch_R, whose objects are "multi-sorted" sets. The most obvious forgetful functors, which send a graph to its set of vertices or its set of edges, are not faithful. However, the functor $V \sqcup E\colon \mathsf{Graph} \to \mathsf{Set}$ considered in Example 1.3.2(ii) that sends a graph to the union of its set of vertices and edges is faithful.

Because faithful functors reflect identifications between parallel morphisms:

LEMMA 1.6.19. *If $U\colon \mathsf{C} \to \mathsf{D}$ is faithful, then any diagram in C whose image commutes in D also commutes in C.*

PROOF. Here a "diagram" in C is a directed graph of morphisms together with a collection of desired composition relations between composable paths of morphisms. If $f_1, \ldots, f_n$ and $g_1, \ldots, g_m$ are parallel sequences of composable morphisms in C so that

$$Uf_n \cdots Uf_1 = Ug_m \cdots Ug_1$$

in D, then by faithfulness (and functoriality) of U, $f_n \cdots f_1 = g_m \cdots g_1$ in C. □

In particular, to prove that a diagram in a concrete category commutes, it suffices to prove commutativity of the induced diagram of underlying sets. This amounts to showing that certain composite functions between underlying sets are the same, and this can be checked by considering the actions of these functions on the elements of their domain.

We close with a word of warning.

REMARK 1.6.20 (on cancellability). We have seen that commutativity of a pair of adjacent squares as in (1.6.10) implies commutativity of the exterior rectangle, but the converse need not hold, as illustrated by the following diagram in Ab, in which the outer rectangle commutes but neither square does:

$$\begin{tikzcd} \mathbb{Z} \arrow[r, "1_{\mathbb{Z}}"] \arrow[d] & \mathbb{Z} \arrow[r] \arrow[d, "1_{\mathbb{Z}}"] & 0 \arrow[d] \\ 0 \arrow[r] & \mathbb{Z} \arrow[r, "1_{\mathbb{Z}}"'] & \mathbb{Z} \end{tikzcd}$$

Neither does commutativity of one of the two squares, plus the outer rectangle, imply commutativity of the other in general. The issue is that a composition relation of the form

$$gh_n \cdots h_1 f = gk_m \cdots k_1 f$$

need not imply that $h_n \cdots h_1 = k_m \cdots k_1$ unless f is "right cancellable" and g is "left cancellable," i.e., unless f is an epimorphism and g is a monomorphism; see Definition 1.2.7.

LEMMA 1.6.21. *Consider morphisms with the indicated sources and targets*

$$\begin{tikzcd} a \arrow[r, "f"] \arrow[d, "g"'] & b \arrow[r, "j"] \arrow[d, "h"] & c \arrow[d, "\ell"] \\ a' \arrow[r, "k"'] & b' \arrow[r, "m"'] & c' \end{tikzcd}$$

and suppose that the outer rectangle commutes. This data defines a commutative rectangle if either:

(i) the right-hand square commutes and m is a monomorphism; or
(ii) the left-hand square commutes and f is an epimorphism.

PROOF. The statements are dual. Assuming (i), $mkg = \ell jf = mhf$ by commutativity of the outer rectangle and right-hand square. Since m is a monomorphism, it follows that $kg = hf$ and thus that the diagram commutes. □

Exercises.

EXERCISE 1.6.i. Show that any map from a terminal object in a category to an initial one is an isomorphism. An object that is both initial and terminal is called a **zero object**.

EXERCISE 1.6.ii. Show that any two terminal objects in a category are connected by a unique isomorphism.

EXERCISE 1.6.iii. Show that any faithful functor reflects monomorphisms. That is, if $F\colon \mathsf{C} \to \mathsf{D}$ is faithful, prove that if Ff is a monomorphism in D, then f is a monomorphism in C. Argue by duality that faithful functors also reflect epimorphisms. Conclude that in any concrete category, any morphism that defines an injection of underlying sets is a monomorphism and any morphism that defines a surjection of underlying sets is an epimorphism.

EXERCISE 1.6.iv. Find an example to show that a faithful functor need not preserve epimorphisms. Argue by duality, or by another counterexample, that a faithful functor need not preserve monomorphisms.

EXERCISE 1.6.v. More specifically, find a concrete category that contains a monomorphism whose underlying function is not injective. Find a concrete category that contains an epimorphism whose underlying function is not surjective. Exercise 4.5.v explains why the latter examples may seem less familiar than the former.

EXERCISE 1.6.vi. A **coalgebra** for an endofunctor $T\colon \mathsf{C} \to \mathsf{C}$ is an object $C \in \mathsf{C}$ equipped with a map $\gamma\colon C \to TC$. A morphism $f\colon (C,\gamma) \to (C',\gamma')$ of coalgebras is a map $f\colon C \to C'$ so that the square

$$\begin{tikzcd} C \arrow[r, "f"] \arrow[d, "\gamma"'] & C' \arrow[d, "\gamma'"] \\ TC \arrow[r, "Tf"'] & TC' \end{tikzcd}$$

commutes. Prove that if (C,γ) is a **terminal coalgebra**, that is a terminal object in the category of coalgebras, then the map $\gamma\colon C \to TC$ is an isomorphism.

1.7. The 2-category of categories

A number of important facts about natural transformations are proven by diagram chasing. In this section, we define "vertical" and "horizontal" composition operations for natural transformations. The upshot is that categories, functors, and natural transformations assemble into a 2-dimensional categorical structure called a *2-category*, a definition that is stated at the conclusion.

In French, a natural transformation is called a *morphisme de foncteurs*. Indeed, for any fixed pair of categories C and D, there is a **functor category** D^{C} whose objects are functors $\mathsf{C} \to \mathsf{D}$ and whose morphisms are natural transformations. Given a functor $F\colon \mathsf{C} \to \mathsf{D}$, its identity natural transformation $1_F\colon F \Rightarrow F$ is the natural transformation whose components $(1_F)_c \coloneqq 1_{Fc}$ are identities. The following lemma defines the composition of morphisms in D^{C}.

LEMMA 1.7.1 (vertical composition). *Suppose $\alpha\colon F \Rightarrow G$ and $\beta\colon G \Rightarrow H$ are natural transformations between parallel functors $F, G, H\colon \mathsf{C} \to \mathsf{D}$. Then there is a natural transformation $\beta \cdot \alpha\colon F \Rightarrow H$ whose components*

$$(\beta \cdot \alpha)_c \coloneqq \beta_c \cdot \alpha_c$$

are defined to be the composites of the components of α and β.

PROOF. Naturality of α and β implies that for any $f\colon c \to c'$ in the domain category, each square, and thus also the composite rectangle, commutes:

$$\begin{tikzcd} Fc \arrow[r, "\alpha_c"] \arrow[d, "Ff"'] & Gc \arrow[r, "\beta_c"] \arrow[d, "Gf"'] & Hc \arrow[d, "Hf"] \\ Fc' \arrow[r, "\alpha_{c'}"'] & Gc' \arrow[r, "\beta_{c'}"'] & Hc' \end{tikzcd}$$ □

COROLLARY 1.7.2. *For any pair of categories C and D, the functors from C to D and natural transformations between them define a category D^{C}.*

PROOF. It remains only to verify that the composition operation defined by Lemma 1.7.1 is associative and unital. It suffices to verify these properties componentwise, and they follow immediately from the associativity and unitality of composition in D. □

REMARK 1.7.3 (sizes of functor categories). Care should be taken with size when discussing functor categories. If C and D are small, then D^{C} is again a small category, but if C and D are locally small, then D^{C} need not be. This is only guaranteed if D is locally small and C is small; see Exercise 1.7.i. In summary, the formation of functor categories defines a bifunctor $\mathsf{Cat}^{\mathrm{op}} \times \mathsf{Cat} \to \mathsf{Cat}$ or $\mathsf{Cat}^{\mathrm{op}} \times \mathsf{CAT} \to \mathsf{CAT}$, but the category of functors between two non-locally small categories may be even larger than these categories are.

The composition operation defined in Lemma 1.7.1 is called **vertical composition**. Drawing the parallel functors horizontally, a composable pair of natural transformations in the category D^{C} fits into a *pasting diagram*

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left=50, "F"] \arrow[r, "G" description] \arrow[r, bend right=50, "H"'] \arrow[r, phantom, bend left=25, "\Downarrow\alpha"] \arrow[r, phantom, bend right=25, "\Downarrow\beta"] & \mathsf{D} \end{tikzcd} \quad = \quad \begin{tikzcd} \mathsf{C} \arrow[r, bend left=50, "F"] \arrow[r, bend right=50, "H"'] \arrow[r, phantom, "\Downarrow\beta\cdot\alpha"] & \mathsf{D} \end{tikzcd}$$

As the terminology suggests, there is also a **horizontal composition** operation

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left=50, "F"] \arrow[r, bend right=50, "G"'] \arrow[r, phantom, "\Downarrow\alpha"] & \mathsf{D} \arrow[r, bend left=50, "H"] \arrow[r, bend right=50, "K"'] \arrow[r, phantom, "\Downarrow\beta"] & \mathsf{E} \end{tikzcd} \quad = \quad \begin{tikzcd} \mathsf{C} \arrow[r, bend left=50, "HF"] \arrow[r, bend right=50, "KG"'] \arrow[r, phantom, "\Downarrow\beta*\alpha"] & \mathsf{E} \end{tikzcd}$$

defined by the following lemma.

LEMMA 1.7.4 (horizontal composition). *Given a pair of natural transformations*

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left=50, "F"] \arrow[r, bend right=50, "G"'] \arrow[r, phantom, "\Downarrow\alpha"] & \mathsf{D} \arrow[r, bend left=50, "H"] \arrow[r, bend right=50, "K"'] \arrow[r, phantom, "\Downarrow\beta"] & \mathsf{E} \end{tikzcd}$$

there is a natural transformation $\beta * \alpha \colon HF \Rightarrow KG$ whose component at $c \in \mathsf{C}$ is defined as the composite of the following commutative square

$$\begin{tikzcd} HFc \arrow[r, "\beta_{Fc}"] \arrow[d, "H\alpha_c"'] \arrow[dr, dashed, "(\beta*\alpha)_c" description] & KFc \arrow[d, "K\alpha_c"] \\ HGc \arrow[r, "\beta_{Gc}"'] & KGc \end{tikzcd} \tag{1.7.5}$$

PROOF. The square (1.7.5) commutes by naturality of $\beta \colon H \Rightarrow K$ applied to the morphism $\alpha_c \colon Fc \to Gc$ in D. To prove that the components $(\beta * \alpha)_c \colon HFc \to KGc$ so-defined are natural, we must show that $KGf \cdot (\beta * \alpha)_c = (\beta * \alpha)_{c'} \cdot HFf$ for any $f \colon c \to c'$ in C. This relation holds on account of the commutative rectangle

$$\begin{tikzcd} HFc \arrow[r, "H\alpha_c"] \arrow[d, "HFf"'] & HGc \arrow[r, "\beta_{Gc}"] \arrow[d, "HGf"'] & KGc \arrow[d, "KGf"] \\ HFc' \arrow[r, "H\alpha_{c'}"'] & HGc' \arrow[r, "\beta_{Gc'}"'] & KGc' \end{tikzcd}$$

The right-hand square commutes by naturality of β. The left-hand square commutes by naturality of α and Lemma 1.6.5, which states that functors, in this case the functor H, preserve commutative diagrams.[42] □

Remark 1.7.6. The natural transformations

$$H\alpha\colon HF \Rightarrow HG, \quad K\alpha\colon KF \Rightarrow KG, \quad \beta F\colon HF \Rightarrow KF, \quad \text{and} \quad \beta G\colon HG \Rightarrow KG$$

appearing in Lemma 1.7.4 are defined by **whiskering** the natural transformations α and β with the functors H and K or F and G, respectively. A precise definition of this construction is given in Exercise 1.7.ii. The terminology is on account of the following graphical depiction of the whiskered composite

$$L\beta F\colon LHF \Rightarrow LKF \qquad \begin{tikzcd} \mathsf{C} \arrow[r, "F"] & \mathsf{D} \arrow[r, bend left, "H"] \arrow[r, bend right, "K"'] \arrow[r, phantom, "\Downarrow\beta"] & \mathsf{E} \arrow[r, "L"] & \mathsf{F} \end{tikzcd}$$

of the natural transformation $\beta\colon H \Rightarrow K$ with the functors F and L. Exercise 1.7.iii explains the particular interest in the case where either L or F is an identity.

Importantly, vertical and horizontal composition can be performed in either order, satisfying the rule of **middle four interchange**:

Lemma 1.7.7 (middle four interchange). *Given functors and natural transformations*

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left=60, "F"] \arrow[r, "G" description] \arrow[r, bend right=60, "H"'] \arrow[r, phantom, shift left=2.5ex, "\Downarrow\alpha"] \arrow[r, phantom, shift right=2.5ex, "\Downarrow\beta"] & \mathsf{D} \arrow[r, bend left=60, "J"] \arrow[r, "K" description] \arrow[r, bend right=60, "L"'] \arrow[r, phantom, shift left=2.5ex, "\Downarrow\gamma"] \arrow[r, phantom, shift right=2.5ex, "\Downarrow\delta"] & \mathsf{E} \end{tikzcd}$$

the natural transformation $JF \Rightarrow LH$ defined by first composing vertically and then composing horizontally equals the natural transformation defined by first composing horizontally and then composing vertically:

$$\begin{tikzcd} \mathsf{C} \arrow[r, bend left=60, "F"] \arrow[r, bend right=60, "H"'] \arrow[r, phantom, "\Downarrow\beta\cdot\alpha"] & \mathsf{D} \arrow[r, bend left=60, "J"] \arrow[r, bend right=60, "L"'] \arrow[r, phantom, "\Downarrow\delta\cdot\gamma"] & \mathsf{E} \end{tikzcd} \quad = \quad \begin{tikzcd} \mathsf{C} \arrow[rr, bend left=40, "JF"] \arrow[rr, "KG" description] \arrow[rr, bend right=40, "LH"'] \arrow[rr, phantom, shift left=2.5ex, "\Downarrow\gamma*\alpha"] \arrow[rr, phantom, shift right=2.5ex, "\Downarrow\delta*\beta"] & & \mathsf{E} \end{tikzcd}$$

Proof. Exercise 1.7.iv. □

Lemmas 1.7.1, 1.7.4, and 1.7.7 prove that categories, functors, and natural transformations assemble into a *2-category*. Aside from this example, we will not meet any other 2-categories in this text. Nonetheless, the following definition is useful as an axiomatization of the composition operations for natural transformations that are available. A succinct introduction to 2-categories and *pasting diagrams*, which are used to display composite natural transformations, can be found in [**KS74**].

Definition 1.7.8. A **2-category** is comprised of:

[42]Naturality of $\beta * \alpha$ could also be deduced from a second commutative rectangle that defines the component $(\beta * \alpha)_c$ as the top-right composite of (1.7.5). The point is that the squares (1.7.5) and the morphisms obtained by applying the four functors HF, HG, KF, and KG to a morphism in C define a commutative cube.

- objects, for example the categories C,
- 1-morphisms between pairs of objects, for example, the functors $\mathsf{C} \xrightarrow{F} \mathsf{D}$, and
- 2-morphisms between parallel pairs of 1-morphisms, for example, the natural transformations $\begin{tikzcd} \mathsf{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "G"'] \arrow[r, phantom, "\Downarrow\alpha"] & \mathsf{D} \end{tikzcd}$

so that:

- The objects and 1-morphisms form a category, with identities $1_{\mathsf{C}} \colon \mathsf{C} \to \mathsf{C}$.
- For each fixed pair of objects C and D, the 1-morphisms $F \colon \mathsf{C} \to \mathsf{D}$ and 2-morphisms between such form a category under an operation called vertical composition, as described in Lemma 1.7.1, with identities $\begin{tikzcd} \mathsf{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "F"'] \arrow[r, phantom, "\Downarrow 1_F"] & \mathsf{D} \end{tikzcd}$.
- There is also a category whose objects are the objects in which a morphism from C to D is a 2-cell $\begin{tikzcd} \mathsf{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "G"'] \arrow[r, phantom, "\Downarrow\alpha"] & \mathsf{D} \end{tikzcd}$ under an operation called horizontal composition, with identities $\begin{tikzcd} \mathsf{C} \arrow[r, bend left, "1_{\mathsf{C}}"] \arrow[r, bend right, "1_{\mathsf{C}}"'] \arrow[r, phantom, "\Downarrow 1_{1_{\mathsf{C}}}"] & \mathsf{C} \end{tikzcd}$. The source and target 1-morphisms of a horizontal composition must have the form described in Lemma 1.7.4.
- The horizontal composite $1_H * 1_F$ of identities for vertical composition must be the identity 1_{HF} for for the composite 1-morphisms.
- The law of middle four interchange described in Lemma 1.7.7 holds.

The reader who has taken the categorical philosophy to heart might ask: What is a morphism between 2-categories? *2-functors* will make an appearance in §4.4.

Exercises.

Exercise 1.7.i. Prove that if C is small and D is locally small, then D^{C} is locally small by defining a monomorphism from the collection of natural transformations between a fixed pair of functors $F, G \colon \mathsf{C} \rightrightarrows \mathsf{D}$ into a set. (Hint: Think about the function that sends a natural transformation to its collection of components.)

Exercise 1.7.ii. Given a natural transformation $\beta \colon H \Rightarrow K$ and functors F and L as displayed in

$$\begin{tikzcd} \mathsf{C} \arrow[r, "F"] & \mathsf{D} \arrow[r, bend left=60, "H"] \arrow[r, bend right=60, "K"'] \arrow[r, phantom, "\Downarrow\beta"] & \mathsf{E} \arrow[r, "L"] & \mathsf{F} \end{tikzcd}$$

define a natural transformation $L\beta F \colon LHF \Rightarrow LKF$ by $(L\beta F)_c = L\beta_{Fc}$. This is the **whiskered composite** of β with L and F. Prove that $L\beta F$ is natural.

Exercise 1.7.iii. Redefine the horizontal composition of natural transformations introduced in Lemma 1.7.4 using vertical composition and whiskering.

Exercise 1.7.iv. Prove Lemma 1.7.7.

EXERCISE 1.7.v. Show that for any category C, the collection of natural endomorphisms of the identity functor 1_{C} defines a commutative monoid, called the **center of the category**. The proof of Proposition 1.4.4 demonstrates that the center of $\mathsf{Ab}_{\mathrm{fg}}$ is the multiplicative monoid $(\mathbb{Z}, \times, 1)$.

EXERCISE 1.7.vi. Suppose the functors and natural isomorphisms

$$\begin{array}{lll} \mathsf{C} \underset{G}{\overset{F}{\rightleftarrows}} \mathsf{D} & \eta\colon 1_{\mathsf{C}} \cong GF & \epsilon\colon FG \cong 1_{\mathsf{D}} \\ \mathsf{D} \underset{G'}{\overset{F'}{\rightleftarrows}} \mathsf{E} & \eta'\colon 1_{\mathsf{D}} \cong G'F' & \epsilon'\colon F'G' \cong 1_{\mathsf{E}} \end{array}$$

define equivalences of categories $\mathsf{C} \simeq \mathsf{D}$ and $\mathsf{D} \simeq \mathsf{E}$. Prove (again) that there is a composite equivalence of categories $\mathsf{C} \simeq \mathsf{E}$ by defining composite natural isomorphisms $1_{\mathsf{C}} \cong GG'F'F$ and $F'FGG' \cong 1_{\mathsf{E}}$.

EXERCISE 1.7.vii. Prove that a bifunctor $F\colon \mathsf{C} \times \mathsf{D} \to \mathsf{E}$ determines and is uniquely determined by:

(i) A functor $F(c, -)\colon \mathsf{D} \to \mathsf{E}$ for each $c \in \mathsf{C}$.
(ii) A natural transformation $F(f, -)\colon F(c, -) \Rightarrow F(c', -)$ for each $f\colon c \to c'$ in C, defined functorially in C.

In other words, prove that there is a bijection between functors $\mathsf{C} \times \mathsf{D} \to \mathsf{E}$ and functors $\mathsf{C} \to \mathsf{E}^{\mathsf{D}}$. By symmetry of the product of categories, these classes of functors are also in bijection with functors $\mathsf{D} \to \mathsf{E}^{\mathsf{C}}$.

CHAPTER 2

Universal Properties, Representability, and the Yoneda Lemma

> . . . a mathematical object X is best thought of in the context of a category surrounding it, and is determined by the network of relations it enjoys with *all* the objects of that category. Moreover, to understand X it might be more germane to deal directly with the functor representing it.
>
> Barry Mazur, "Thinking about Grothendieck" [**Maz16**]

The aim in this chapter is to explain what it means to say that the natural numbers is the universal discrete dynamical system, that the Sierpinski space is the universal space with an open subset, or that the complete graph on n-vertices is the universal n-colored graph. Universal properties expressed in this plain language manner might appear somewhat ad hoc, but this is a false impression. By the chapter's end, we will see several equivalent ways in which the notion of universal property can be made precise. The key input to the theory developed here is something that is a priori non-obvious: a functor valued in the category of sets. The category of sets plays a special role because traditional approaches to the foundations of mathematics are based on set theory, with a wide variety of mathematical objects defined to be sets with additional structure.

To illustrate, consider the set of vertex colorings of a graph subject to the requirement that adjacent vertices are assigned distinct colors. There is a contravariant functor from the category of graphs to the category of sets that takes a graph to the set of n-colorings of its vertices. To explain the contravariance, note that an n-coloring of a graph G and a graph homomorphism $G' \to G$ induce an n-coloring of G' that colors each vertex of G' to match the color of its image: as graph homomorphisms preserve the incidence relation on vertices, any two adjacent vertices of G' are assigned distinct colors. This defines the action of the functor $n\text{-Color} \colon \mathsf{Graph}^{\mathrm{op}} \to \mathsf{Set}$ on morphisms.[1]

The graph with the fewest vertices and fewest edges that can be colored with no less than n colors is the complete graph on n vertices K_n. Indeed, the functor n-Color encodes a *universal property* of the graph K_n in the sense that the graph K_n *represents* the functor n-Color. Specifically, this means that there is a natural bijection between the set $n\text{-Color}(G)$ of n-colorings of a graph G and the set of graph homomorphisms $G \to K_n$. In §2.1, we introduce the notion of a *representable functor* along with a plethora of examples, intended to convey the ubiquity of this abstract notion.

[1]The **four color theorem** states that if G is a simple planar graph, then the set $4\text{-Color}(G)$ is non-empty. Functoriality implies that any graph admitting a graph homomorphism to a planar graph also admits a 4-coloring. For instance, complete bipartite graphs, which are typically non-planar, will admit homomorphisms of this type.

By the Yoneda lemma, which we prove in §2.2, this universal property uniquely characterizes the graph K_n. The Yoneda lemma also establishes a correspondence between natural endomorphisms of the functor n-Color—"color permutations"—and symmetries of the graph K_n and proves that there is no uniform way to convert an m-coloring into an n-coloring if $m > n$, this result deduced from the fact that there are no graph homomorphisms $K_m \to K_n$.

The Yoneda lemma is arguably the most important result in category theory, although it takes some time to explore the depths of the consequences of this simple statement. In §2.3, we define the notion of *universal element* that witnesses a universal property of some object in a locally small category. The universal element witnessing the universal property of the complete graph is an n-coloring of K_n, an element of the set $n\text{-Color}(K_n)$. In §2.4, we use the Yoneda lemma to show that the pair comprised of an object characterized by a universal property and its universal element defines either an initial or a terminal object in the *category of elements* of the functor that it represents. This gives precise meaning to the term **universal**: it is a synonym for either "initial" or "terminal," with context disambiguating between the two cases. For instance, K_n is the **terminal n-colored graph**: the terminal object in the category of n-colored graphs and graph homomorphisms that preserve the coloring of vertices.

2.1. Representable functors

> The further you go in mathematics, especially pure mathematics, the more universal properties you will meet.
>
> Tom Leinster, *Basic Category Theory* [**Lei14**]

The most basic formulation of a universal property is to say that a particular object defines an initial or terminal object in its ambient category. The problem with this paradigm is that the most familiar categories—for instance, of sets, spaces, groups, modules, and so on—tend to have uninteresting initial and terminal objects. To express the universal properties of more complicated objects, one has to cook up a less familiar category. For example:

Example 2.1.1. A set X with an endomorphism $f\colon X \to X$ and a distinguished element x_0 is called a **discrete dynamical system**. This data allows one to consider the discrete-time evolution of the initial element x_0, a sequence defined by $x_{n+1} := f(x_n)$. The principle of mathematical recursion asserts that the natural numbers $\mathbb{N}$, the successor function $s\colon \mathbb{N} \to \mathbb{N}$, and the element $0 \in \mathbb{N}$ define the **universal discrete dynamical system**: which is to say, there is a unique function $r\colon \mathbb{N} \to X$ so that $r(n) = x_n$ for each n, i.e., so that $r(0) = x_0$ and so that the diagram

$$\begin{tikzcd} \mathbb{N} \arrow[r, "s"] \arrow[d, "r"'] & \mathbb{N} \arrow[d, "r"] \\ X \arrow[r, "f"'] & X \end{tikzcd} \tag{2.1.2}$$

commutes.

In this section, we introduce a vehicle for expressing universal properties that is less-contrived than spontaneously generating an appropriate category in which the universal object is initial or terminal. The notion that we introduce can be understood as a generalization of the universal properties describing initial or terminal objects. In §2.4, we will show that it is not a true generalization.

To say that an object $c \in \mathsf{C}$ is initial is to say that for all other objects x, the set $\mathsf{C}(c, x)$ is a singleton. Dually, $c \in \mathsf{C}$ is terminal, if and only if for all $x \in \mathsf{C}$ the set $\mathsf{C}(x, c)$ is a singleton. These properties can be expressed more categorically as characterizations of the co- and contravariant functors $\mathsf{C}(c, -)$ and $\mathsf{C}(-, c)$ represented by c that were introduced in Definition 1.3.11:

$$\begin{array}{ccc} \mathsf{C} & \xrightarrow{\mathsf{C}(c,-)} & \mathsf{Set} \\ x & \mapsto & \mathsf{C}(c, x) \\ f\big\downarrow & \mapsto & \big\downarrow f_* \\ y & \mapsto & \mathsf{C}(c, y) \end{array} \qquad \begin{array}{ccc} \mathsf{C}^{\mathrm{op}} & \xrightarrow{\mathsf{C}(-,c)} & \mathsf{Set} \\ x & \mapsto & \mathsf{C}(x, c) \\ f\big\downarrow & \mapsto & \big\uparrow f^* \\ y & \mapsto & \mathsf{C}(y, c) \end{array}$$

DEFINITION 2.1.3 (a representable characterization of initial or terminal objects).

(i) An object c in a category C is **initial** if and only if the functor $\mathsf{C}(c, -) : \mathsf{C} \to \mathsf{Set}$ is naturally isomorphic to the constant functor $* : \mathsf{C} \to \mathsf{Set}$ that sends every object to the singleton set.

(ii) An object $c \in \mathsf{C}$ is **terminal** if and only if the functor $\mathsf{C}(-, c) : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ is naturally isomorphic to the constant functor $* : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ that sends every object to the singleton set.

In terminology we now introduce, Definition 2.1.3 asserts that C has an initial object if and only if the constant functor $* : \mathsf{C} \to \mathsf{Set}$ is representable, and dually, C has a terminal object if and only if $* : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ is representable.

DEFINITION 2.1.4.

(i) A covariant or contravariant functor F from a locally small category C to Set is **representable** if there is an object $c \in \mathsf{C}$ and a natural isomorphism between F and the functor of appropriate variance[2] represented by c, in which case one says that the functor F is **represented by** the object c.

(ii) A **representation** for a functor F is a choice of object $c \in \mathsf{C}$ together with a specified natural isomorphism $\mathsf{C}(c, -) \cong F$, if F is covariant, or $\mathsf{C}(-, c) \cong F$, if F is contravariant.

The domain C of a representable functor is required to be locally small so that the hom-functors $\mathsf{C}(c, -)$ and $\mathsf{C}(-, c)$ are valued in the category of sets.

As in the special case of Definition 2.1.3, a representable functor $F : \mathsf{C} \to \mathsf{Set}$ or $F : \mathsf{C}^{\mathrm{op}} \to \mathsf{Set}$ encodes a **universal property** of its representing object. Put colloquially, a universal property of an object X is a description of the covariant functor $\mathrm{Hom}(X, -)$ or of the contravariant functor $\mathrm{Hom}(-, X)$ associated to that object. Many examples occur "in nature."[3]

EXAMPLE 2.1.5. The following covariant functors are representable.

(i) The identity functor $1_{\mathsf{Set}} : \mathsf{Set} \to \mathsf{Set}$ is represented by the singleton set 1. That is, for any set X, there is a natural isomorphism $\mathsf{Set}(1, X) \cong X$ that defines a bijection

[2]Some use the term "corepresentable" for covariant representable functors, reserving "representable" for the contravariant case. We argue that this distinction is unnecessary since the variance of the functor F ought to be evident from its definition.

[3]If a friend or oracle is nearby, we suggest first encountering the examples listed in 2.1.5 and 2.1.6 as exercises. Have the friend list the functors that are mentioned and challenge yourself to find the representing objects.

between elements $x \in X$ and functions $x\colon 1 \to X$ carrying the singleton element to x. Naturality says that for any $f\colon X \to Y$, the diagram

$$\begin{tikzcd} \mathsf{Set}(1,X) \arrow[r, "\cong"] \arrow[d, "f_*"'] & X \arrow[d, "f"] \\ \mathsf{Set}(1,Y) \arrow[r, "\cong"'] & Y \end{tikzcd}$$

commutes, i.e., that the composite function $1 \xrightarrow{x} X \xrightarrow{f} Y$ corresponds to the element $f(x) \in Y$, as is evidently the case.

(ii) The forgetful functor $U\colon \mathsf{Group} \to \mathsf{Set}$ is represented by the group $\mathbb{Z}$. That is, for any group G, there is a natural isomorphism $\mathsf{Group}(\mathbb{Z}, G) \cong UG$ that associates, to every element $g \in UG$, the unique homomorphism $\mathbb{Z} \to G$ that maps the integer 1 to g. This defines a bijection because every homomorphism $\mathbb{Z} \to G$ is determined by the image of the generator 1; that is to say, $\mathbb{Z}$ is the **free group on a single generator**. This bijection is natural because the composite group homomorphism $\mathbb{Z} \xrightarrow{g} G \xrightarrow{\phi} H$ carries the integer 1 to $\phi(g) \in H$.

(iii) For any unital ring R, the forgetful functor $U\colon \mathsf{Mod}_R \to \mathsf{Set}$ is represented by the R-module R. That is, there is a natural bijection between R-module homomorphisms $R \to M$ and elements of the underlying set of M, in which $m \in UM$ is associated to the unique R-module homomorphism that carries the multiplicative identity of R to m; this is to say, R is the **free R-module on a single generator**. This explains the appearance of the abelian group $\mathbb{Z}$ and the vector space $\Bbbk$ in Definition 1.6.3, where maps with these domains were used to specify elements in the codomains.

(iv) The functor $U\colon \mathsf{Ring} \to \mathsf{Set}$ is represented by the unital ring $\mathbb{Z}[x]$, the polynomial ring in one variable with integer coefficients. A unital ring homomorphism $\mathbb{Z}[x] \to R$ is uniquely determined by the image of x; put another way, $\mathbb{Z}[x]$ is the **free unital ring on a single generator**.

(v) The functor $U(-)^n\colon \mathsf{Group} \to \mathsf{Set}$ that sends a group G to the set of n-tuples of elements of G is represented by the **free group F_n on n generators**. Similarly, the functor $U(-)^n\colon \mathsf{Ab} \to \mathsf{Set}$ is represented by the **free abelian group $\bigoplus_n \mathbb{Z}$ on n generators**.

(vi) More generally, any group presentation, such as

$$S_3 := \left\langle s, t \;\middle|\; s^2 = t^2 = 1, sts = tst \right\rangle,$$

defines a functor $\mathsf{Group} \to \mathsf{Set}$ that carries a group G to the set

$$\left\{(g_1, g_2) \in G^2 \;\middle|\; g_1^2 = g_2^2 = e, g_1g_2g_1 = g_2g_1g_2\right\}.$$

The functor is represented by the group admitting the given presentation, in this case by the symmetric group S_3 on three elements: the presentation tells us that homomorphisms $S_3 \to G$ are classified by pairs of elements $g_1, g_2 \in G$ satisfying the listed relations.

(vii) The functor $(-)^\times\colon \mathsf{Ring} \to \mathsf{Set}$ that sends a unital ring to its set of units is represented by the ring $\mathbb{Z}[x, x^{-1}]$ of Laurent polynomials in one variable. That is to say, a ring homomorphism $\mathbb{Z}[x, x^{-1}] \to R$ may be defined by sending x to any unit of R and is completely determined by this assignment, and moreover there are no ring homomorphisms that carry x to a non-unit.

(viii) The forgetful functor $U: \mathsf{Top} \to \mathsf{Set}$ is represented by the singleton space: there is a natural bijection between elements of a topological space and continuous functions from the one-point space.

(ix) The functor $\mathrm{ob}: \mathsf{Cat} \to \mathsf{Set}$ that takes a small category to its set of objects is represented by the terminal category $\mathbb{1}$: a functor $\mathbb{1} \to \mathsf{C}$ is no more and no less than a choice of object in C.

(x) The functor $\mathrm{mor}: \mathsf{Cat} \to \mathsf{Set}$ that takes a small category to its set of morphisms is represented by the category $\mathbb{2}$: a functor $\mathbb{2} \to \mathsf{C}$ is no more and no less than a choice of morphism in C. In this sense, the category $\mathbb{2}$ is the **free** or **walking arrow**.

(xi) The functor $\mathrm{iso}: \mathsf{Cat} \to \mathsf{Set}$ that takes a small category to its set of isomorphisms (pointing in a specified direction) is represented by the category $\mathbb{I}$, with two objects and exactly one morphism in each hom-set. In this sense, the category $\mathbb{I}$ is the **free** or **walking isomorphism**.

(xii) The functor $\mathrm{comp}: \mathsf{Cat} \to \mathsf{Set}$ that takes a small category to the set of composable pairs of morphisms in it is represented by the category $\mathbb{3}$. Generalizing, the ordinal $\mathbb{n} + \mathbb{1} = 0 \to 1 \to \cdots \to n$ represents the functor that takes a small category to the set of paths of n composable morphisms in it.

(xiii) The forgetful functor $U: \mathsf{Set}_* \to \mathsf{Set}$ is represented by the two-element based set: based functions out of this set correspond naturally and bijectively to elements of the target based set, the element in question being the image of the non-basepoint element.

(xiv) The functor $\mathrm{Path}: \mathsf{Top} \to \mathsf{Set}$ that carries a topological space to its set of paths and the functor $\mathrm{Loop}: \mathsf{Top}_* \to \mathsf{Set}$ that carries a based space to its set of based loops are each representable by definition, by the unit interval I and the based circle S^1, respectively. A **path** in X is a continuous function $I \to X$ while a (based) **loop** in X is a based continuous function $S^1 \to X$.

The adjective "free" is reserved for universal properties expressed by covariant represented functors. It could be applied to any of the objects listed in Example 2.1.5: $\mathbb{2}$ is the free category with an arrow, S^1 is the free space containing a loop. The dual term "cofree," for universal properties expressed by contravariant represented functors, is less commonly used.

EXAMPLE 2.1.6. The following contravariant functors are representable.

(i) The contravariant power set functor $P: \mathsf{Set}^{\mathrm{op}} \to \mathsf{Set}$ is represented by the set $\Omega = \{\top, \bot\}$ with two elements. The natural isomorphism $\mathsf{Set}(A, \Omega) \cong PA$ is defined by the bijection that associates a function $A \to \Omega$ with the subset that is the preimage of $\top$; reversing perspectives, a subset $A' \subset A$ is identified with its **classifying function** $\chi_{A'}: A \to \Omega$, which sends exactly the elements of A' to the element $\top$. The naturality condition stipulates that for any function $f: A \to B$, the diagram

$$\begin{tikzcd} \mathsf{Set}(B,\Omega) \arrow[r, "\cong"] \arrow[d, "f^*"'] & PB \arrow[d, "f^{-1}"] \\ \mathsf{Set}(A,\Omega) \arrow[r, "\cong"'] & PA \end{tikzcd}$$

commutes. That is, naturality asserts that given a function $\chi_{B'}: B \to \Omega$ classifying the subset $B' \subset B$, the composite function $A \xrightarrow{f} B \xrightarrow{\chi_{B'}} \Omega$ classifies the subset $f^{-1}(B') \subset A$.

(ii) The functor $O\colon \mathsf{Top}^{\mathrm{op}} \to \mathsf{Set}$ that sends a space to its set of open subsets is represented by the **Sierpinski space** S, the topological space with two points, one closed and one open. The natural bijection $\mathsf{Top}(X, S) \cong O(X)$ associates a continuous function $X \to S$ to the preimage of the open point. This bijection is natural because a composite function $Y \to X \to S$ classifies the preimage of the open subset of X under the function $Y \to X$.

(iii) The Sierpinski space also represents the functor $C\colon \mathsf{Top}^{\mathrm{op}} \to \mathsf{Set}$ that sends a space to its set of closed subsets. Composing the natural isomorphisms $O \cong \mathsf{Top}(-, S) \cong C$ we see that the closed set and open set functors are naturally isomorphic. The composite natural isomorphism carries an open subset to its complement, which is closed. This recovers the natural isomorphism described in Example 1.4.3(v).

(iv) The functor $\mathrm{Hom}(- \times A, B)\colon \mathsf{Set}^{\mathrm{op}} \to \mathsf{Set}$ that sends a set X to the set of functions $X \times A \to B$ is represented by the set B^A of functions from A to B. That is, there is a natural bijection between functions $X \times A \to B$ and functions $X \to B^A$. This natural isomorphism is referred to as **currying** in computer science; by fixing a variable in a two-variable function, one obtains a family of functions in a single variable.

(v) The functor $U(-)^*\colon \mathsf{Vect}_{\Bbbk}^{\mathrm{op}} \to \mathsf{Set}$ that sends a vector space to the set of vectors in its dual space is represented by the vector space $\Bbbk$, i.e., linear maps $V \to \Bbbk$ are, by definition, precisely the vectors in the dual space V^*.

(vi) For any fixed abelian group A and any $n \geq 0$, *singular cohomology with coefficients in A* defines a functor $H^n(-; A)\colon \mathsf{Top}^{\mathrm{op}} \to \mathsf{Ab}$. As in Example 1.3.2(iii), this functor is a homotopy invariant, factoring through the quotient $\mathsf{Top} \to \mathsf{Htpy}$ to define a functor $H^n(-; A)\colon \mathsf{Htpy}^{\mathrm{op}} \to \mathsf{Ab}$. Passing to underlying sets and restricting to a subcategory of "nice" spaces, such as the *CW complexes*, the resulting functor $H^n(-; A)\colon \mathsf{Htpy}_{\mathrm{CW}}^{\mathrm{op}} \to \mathsf{Set}$ is represented by the **Eilenberg–MacLane space** $K(A, n)$. That is, for any CW complex X, homotopy classes of maps $X \to K(A, n)$ stand in bijection with elements of the nth singular cohomology group $H^n(X; A)$ of X with coefficients in A.

(vii) A **classifying space** for a topological group G is a CW complex BG that represents the functor $\mathsf{Htpy}_{\mathrm{CW}}^{\mathrm{op}} \to \mathsf{Set}$ that takes a CW complex to the set of isomorphism classes of *principal G-bundles* over it.[4]

Our language asserts that a representation encodes some sort of universal property of its representing object, but many questions remain:

- How unique are these universal properties? If two objects represent the same functor, are they isomorphic?
- What data is involved in the construction of a natural isomorphism between a representable functor F and the functor represented by an object c?
- How do the universal properties expressed by representable functors relate to initial and terminal objects—our first paradigm for "universality"?

Answers to all of these questions will make use of the Yoneda lemma, to which we now turn.

Exercises.

[4]If G is a discrete group, the CW complex BG is built from a collection of combinatorial data that defines the *nerve* of the category $\mathsf{B}G$ (see Exercise 6.5.iv), hence our use of this notation for 1-object groupoids.