This ever-expanding document started as a set of study notes and exercises and gradually outgrew itself to become more encyclopedic. Having all these notes in one place is quite helpful for both expressing my own thoughts clearly and for later reference. It is also helpful for tracking connections between seemingly unrelated concepts — the entire document is inter-hyperlinked. Even though I aim at understanding every concept in the way it is meant to be used, most concepts are presented insomuch as they are relevant to probability and optimization.

Since these are study notes, they will naturally have a lot of errors, so read them with caution. The document claims no referential nor pedagogical value. Everything is written at the level of abstraction I am comfortable with. Furthermore, some sections in the document are much more polished than others. Feel free to contact me if something in this document happens to distress you.

I tried putting citations on as many things as possible. The citations themselves are usually put in the left margin. If there is no citation on a definition or theorem, that means that I have either recalled it from memory or discovered it on my own. Many of the unoriginal definitions and theorems are restated. The simple proofs are mostly original, and the difficult ones are, often loosely, based on proofs from the places cited. Some proofs simply state “Trivial.” in order to distinguish themselves from the proofs that are omitted for other reasons.

See the remark in the introduction of the sections on mathematical logic and set theory for clarification regarding seemingly arbitrary conventions.
# Contents

## 1. Numbers
1.1. Natural numbers ........................................ 7
1.2. Integers .................................................. 18
1.3. Real numbers ............................................. 25
1.4. Complex numbers ........................................ 27

## 2. Real analysis
2.1. Topology of Euclidean spaces ............................ 30
2.2. Real-valued functions .................................... 33
2.3. Real convergence ........................................ 34
2.4. Real differentiability ..................................... 37
2.5. Real series ............................................... 38
2.6. Riemann integration ...................................... 42
2.7. Line integrals ............................................. 50
2.8. Total variation ............................................ 52
2.9. Nonsmooth derivatives ................................... 53
2.10. Convex functions ........................................ 54
2.11. Subdifferentials .......................................... 61
2.12. Clarke generalized gradients ......................... 62

## 3. Complex analysis
3.1. Complex functions ........................................ 64
3.2. Series ..................................................... 65
3.3. Power series ............................................... 73
3.4. Trigonometric functions .................................. 76
3.5. Exponential function ...................................... 82
3.6. Trigonometric polynomials ............................... 88
3.7. Special functions ......................................... 89
3.8. Norms ..................................................... 90

## 4. Functional analysis
4.1. Topological groups ....................................... 97
4.2. Topological vector spaces ................................. 98
4.3. The Hahn-Banach theorem ................................ 100
4.4. Frechet spaces ........................................... 101
4.5. Banach spaces ............................................ 102
4.6. Hilbert spaces ............................................ 104
4.7. Asplund spaces ........................................... 105
4.8. Minkowski functionals ................................... 106
4.9. Dentable sets ............................................. 107
5. **Approximation theory**  
5.1. Lagrange polynomials ........................................... 123  
5.2. Bernstein inequalities ........................................... 125

6. **General topology**  
6.1. Topological spaces ............................................. 129  
6.2. Topological nets .................................................. 140  
6.3. Function convergence ........................................... 145  
6.4. Topological continuity ......................................... 147  
6.5. Initial and final topologies .................................... 150  
6.6. Separation axioms ............................................... 153  
6.7. Connected spaces ............................................... 155  
6.8. Compact spaces .................................................. 156  
6.9. Baire spaces ...................................................... 158  
6.10. Uniform spaces .................................................. 160

7. **Metric spaces**  
7.1. Metric topology .................................................... 169  
7.2. Complete metric spaces ........................................ 174  
7.3. Hausdorff distance ............................................... 179  
7.4. Totally bounded sets ............................................ 181  
7.5. Noncompactness measures ..................................... 184  
7.6. Lipschitz continuity .............................................. 187  
7.7. Function oscillation ............................................. 189

8. **Geometry**  
8.1. Affine coordinate systems .................................... 191  
8.2. Vector space geometry .......................................... 195  
8.3. Analytic geometry in the plane ............................... 201  
8.4. Manifolds .......................................................... 207  
8.5. Quadratic curves ................................................. 208

9. **Group theory**  
9.1. Monoids ............................................................ 213  
9.2. Groups .............................................................. 221  
9.3. Free groups ......................................................... 231  
9.4. Group actions ..................................................... 237  
9.5. Abelian groups ..................................................... 246

10. **Ring theory**  
10.1. Semirings ......................................................... 254  
10.2. Semimodules ....................................................... 261
10.3. Semiring ideals ................................................. 269
10.4. Algebras over semirings .................................... 277
10.5. Rings ....................................................... 284
10.6. Modules .................................................. 295
10.7. Univariate polynomials ..................................... 308
10.8. Algebras over rings ......................................... 316
10.9. Integral domains ............................................ 324
10.10. Field extensions ............................................ 340

11. Linear algebra .................................................. 346
11.1. Matrices over rings ......................................... 347
11.2. Matrices over fields ........................................ 361
11.3. Bilinear forms ............................................. 369
11.4. Algebraic dual spaces ...................................... 374
11.5. Diagonalization ............................................ 377

12. Mathematical logic ............................................. 378
12.1. Formal languages .......................................... 382
12.2. Boolean functions ........................................ 393
12.3. Propositional logic ....................................... 396
12.4. First-order logic .......................................... 406
12.5. First-order satisfiability .................................. 415
12.6. First-order models ........................................ 427
12.7. Deductive systems ........................................ 438
12.8. Logical theories .......................................... 453

13. Set theory ....................................................... 459
13.1. Naïve set theory ........................................... 461
13.2. Relations .................................................. 473
13.3. Functions .................................................. 482
13.4. Zermelo-Fraenkel set theory ............................. 497
13.5. Well-ordered sets ......................................... 508
13.6. Ordinals .................................................... 512
13.7. Cardinals ................................................... 528
13.8. Transfinite arithmetic .................................... 540
13.9. Von Neumann's cumulative hierarchy ................. 552
13.10. Grothendieck universes ................................ 560

14. Category theory ................................................. 565
14.1. Categories .................................................. 566
14.2. Functors ................................................... 576
14.3. Category equivalences .................................... 594
14.4. Category adjunctions ..................................... 605
14.5. Categorical limits ........................................ 620
1. Numbers

Numbers are perhaps the most ubiquitous concept in mathematics. Even among non-mathematicians, division by zero or $0.999 \ldots = 1$ seem to be a common topic of discourse, either as a joke or a sincere misunderstanding.

The aforementioned topics were studied extensively by mathematicians and have simple justifications from the point of view of abstract mathematics:

- We may want to somehow define division by zero in the field $\mathbb{R}$ of real numbers, however that would introduce zero divisors and hence deprive $\mathbb{R}$ of being an integral domain. The cancellative property of multiplication would not hold, and hence $xy = zy$ would not imply that $x = z$ when $y \neq 0$.

Our familiar arithmetic of real numbers heavily relies on the cancellative property, therefore we simply disallow division by zero.

- The set $\mathbb{R}$ of real numbers is a uniform space and thus every fundamental sequence is convergent by theorem 353 (Cauchy’s net convergence criterion). Furthermore, since $\mathbb{R}$ is also a Hausdorff space, by proposition 308, every fundamental sequence has a unique limit.

Now consider the following two fundamental sequences:

$1, 1, 1, 1, \ldots,$
$0, 0.9, 0.99, \ldots.$

Their absolute difference

$1, 0.1, 0.01, \ldots$

converges to 0.

Therefore, the two original sequences converge to the same real number, namely 1.

Unfortunately, a formal study of numbers also leads to artifacts such as the nonstandard natural numbers discussed in remark 4.

We will describe some basic properties of the common number systems:

- The set $\mathbb{N}$ of natural numbers from the perspective of section 12 (Mathematical logic).
- The set $\mathbb{Z}$ of integers from the perspective of section 10 (Ring theory).
- The set $\mathbb{Q}$ of rational numbers, only briefly mentioned.
- The set $\mathbb{R}$ of real numbers from the perspective of section 2 (Real analysis).
- The set $\mathbb{C}$ of complex numbers from the perspective of section 3 (Complex analysis) and section 10 (Ring theory).
1.1. Natural numbers

Definition 1. Peano arithmetic (commonly abbreviated as PA) is a theory in first-order predicate logic for describing natural numbers and their operations. It can also be formulated in second-order logic or entirely within set theory (especially considering that we are working inside an ambient standard transitive model of ZFC+U), however in this document we usually give preference to the first-order logic formulation of a theory.

Peano’s original axioms referred to sets rather than first-order logic, and we actually work in this document, however we prefer to use a modern formulation of Peano arithmetic.

The language of the theory consists of

(a) A constant 0 for representing zero. We can alternatively require a constant for 1, but this would lead to worse metamathematical properties as discussed in remark 2.

(b) A unary functional symbol \( s \), called the successor operation.

The successor function is only a technicality used for establishing basic properties and for defining addition and multiplication, both in this subsection and in definition 980.

We will only use the abstract successor operation prior to proving the familiar properties of addition and multiplication, although we will later use the ordinal successor operation for building a model of PA — see theorem 981.

(c) An infix binary functional symbol \( + \) for denoting addition.

See proposition 6 for the algebraic properties of natural number addition.

(d) Another infix binary functional symbol \( \cdot \) for denoting multiplication. Outside the object language we usually use juxtaposition instead.

See proposition 8 for the algebraic properties of natural number multiplication.

As with all semirings, multiplication has higher priority than addition. In the unambiguous language defined in example 786, this means that we can use the shorthand \( \xi + \eta \cdot \zeta \) for \( ((\xi \cdot \eta) + \zeta) \). We often use juxtaposition for denoting multiplication.

As usual, in order to avoid parentheses, we assume that multiplication has a higher precedence and thus the right-hand side of axiom (PA7) should be parenthesized as \((\eta \cdot \xi) + \xi\). We avoid excessive parentheses in formulas as per our convention remark 808.

We impose the following base axioms:

PA1 The successor function is injective. This can be stated as follows:

\[
s(\xi) \neq s(\eta) \rightarrow \xi \neq \eta. \tag{PA1}
\]

We use here the convention for implicit universal quantification described in remark 781 (g).

PA2 Zero is not the successor of any natural number. Symbolically,

\[
\neg \exists \xi \cdot (s(\xi) \neq 0). \tag{PA2}
\]
The **axiom schema of induction** roughly states that for a property to hold for all natural numbers it is sufficient for the following two conditions to be met:

- The property holds for 0.
- We can prove that it holds for any number by assuming that it holds for its predecessor.

See the proof of proposition 5 for a detailed discussion.

To describe this formally, we state that for any variables $\xi$ and $\eta$ and any formula $\varphi$ not containing $\eta$ as a free variable, the following is an axiom:

\[
\begin{aligned}
\quad &\varphi[\xi \mapsto 0] \land \forall \eta \cdot \left( \varphi[\xi \mapsto \eta] \rightarrow \varphi[\xi \mapsto s(\eta)] \right) \rightarrow \forall \eta \cdot \varphi[\xi \mapsto \eta]. \\
\text{base case} &\quad \text{inductive hypothesis} &\quad \text{inductive step} &\quad \text{conclusion}
\end{aligned}
\] (PA3)

It is important to highlight that $\varphi$ may have any set of free variables, as long as $\eta$ is not among them. As explained in remark 781 (g), we avoid excessive universal quantification. Of course, the axiom is only interesting if $\xi \in \text{Free}(\varphi)$. If $\zeta_1, \ldots, \zeta_n$ are all the other free variables of $\varphi$, then the universal closure of the corresponding axiom is

\[
\begin{aligned}
\forall \zeta_1 \cdot \ldots \cdot \forall \zeta_n, &\quad \left( \varphi[\xi \mapsto 0] \land \forall \eta \cdot \left( \varphi[\xi \mapsto \eta] \rightarrow \varphi[\xi \mapsto s(\eta)] \right) \rightarrow \forall \eta \cdot \varphi[\xi \mapsto \eta] \right).
\end{aligned}
\] (PA3’)

Thus, the axiom holds for any assignment for the variables $\zeta_1, \ldots, \zeta_n$. For this reason, we call these variables **parameters**. Parameters in axiom schemas are further discussed in definition 928 in relation to comprehension in set theory.

See remark 1027 for a more detailed discussion of induction in general and theorem 979 (Recursion theorem) for the corresponding recursion principle.

The theory we obtain without the binary operations and with only the axioms (PA1)-(PA3) is itself sometimes called Peano arithmetic. The operations are defined inductively, however, and there is no way for us to formalize them within the object theory without adding them to the language and theory itself.

**PA4+5** The next two axioms inductively define how addition is supposed to work:

\[
\begin{aligned}
\xi + 0 &\equiv \xi \quad \text{(PA4)} \\
\xi + s(\eta) &\equiv s(\xi + \eta) \quad \text{(PA5)}
\end{aligned}
\]

**PA6+7** The final two axioms are for multiplication:

\[
\begin{aligned}
\xi \cdot 0 &\equiv 0 \quad \text{(PA6)} \\
\xi \cdot s(\eta) &\equiv \xi \cdot \eta + \xi \quad \text{(PA7)}
\end{aligned}
\]
Remark 2. It is common to consider the first natural numbers to be 0 (e.g. [Aut20, exmpl. 17.6]). Peano himself, however, considered 1 to be the first natural number - see [Pea89, p. 1].

Whether 0 is considered to be a natural number is a matter of convention. The operations defined via (PA4)-(PA7) can be modified to work if 1 was instead the first natural number.

We make choose for \( \mathbb{N} \) to start with 0, however we often avoid referring to the set \( \mathbb{N} \) of natural numbers and instead rely on the concepts “nonnegative” and “positive” integers formally defined in definition 17.

**Definition 3.** We define the set of natural numbers \( \mathbb{N} \) as the smallest inductive set \( \omega \) with the interpretation described in theorem 981.

We do not depend on any particular properties of \( \omega \), but we use it because our construction of it is careful and purposely does not use natural numbers to avoid circularity. We are working in an ambient standard transitive model of ZFC+U and hence we will conflate \( \mathbb{N} \) with \( \omega \) as sets, however the first is also a structure of first-order logic.

We use the usual notation

\[
0 := \emptyset \\
1 := \text{succ}(\emptyset) = \{\emptyset\} \\
2 := \text{succ}(\text{succ}(\emptyset)) = \{\emptyset, \{\emptyset\}\} \\
\vdots
\]

and continue to use the notation functional symbols from definition 1, however we now denote the corresponding interpretations in the structure \( \mathbb{N} \).

See example 786 for a simple grammar that produces numeric symbols in their decimal notation.

**Remark 4.** At this point, we have two kinds of natural numbers:

- We have natural numbers within the metatheory. This is our mental model of the natural numbers, and it is used for distinguishing between “unary” functional symbols like \( s \) and “binary” functional symbols like +. This is mostly used within logic itself.

- We have the set of natural numbers \( \mathbb{N} \) defined in definition 3. These are the numbers which we have defined formally, whose properties we study and the numbers which we use in the entire document. The properties of \( \mathbb{N} \) help us develop a better mental model, which in turn changes our perception of the natural numbers within the metatheory.

We want the two sets of natural numbers to coincide. This is important when talking about, for example, sequences. If a number in \( \mathbb{N} \) is not a natural number within the metatheory, we say that it is nonstandard. The existence of nonstandard models is guaranteed by theorem 912 (Upward Löwenheim-Skolem theorem). There cannot be numbers in the metatheory that are not in \( \mathbb{N} \) because a model of PA cannot have a finite domain and the natural numbers are the smallest metalogical infinite set.

A model of PA which contains precisely the numbers in the metatheory is called a standard model. For the purpose of this document, it is sufficient to accept the convention that \( \mathbb{N} \) is a standard model of PA.

---

1 bg: естествени числа, ru: натуральные числа
Proposition 5. Every nonzero natural number has a unique predecessor. More precisely, zero has no predecessor and for any nonzero number \( n \) there exists a unique number \( m \) such that \( n = s(m) \). We will denote this predecessor by \( p(n) \).

Proof. This proof is exemplar because it clearly demonstrates both the distinction between inductive and deductive reasoning and the role of the main three axioms.

- The axiom (PA1) states that the function \( s \) is injective. By the equivalences in definition 971 (a) its inverse multi-valued function is actually a single-valued partial function. Denote this inverse by \( p \).

- The axiom (PA2) states that the function \( s \) is not surjective. By the equivalences in definition 971 (b), the inverse \( p \) is not a total function.

What we have shown up until this point in the proof is deductive — we have restated the first two axioms of PA and used some equivalent conditions that allowed us to deduce properties of the inverse function \( p \) of \( s \). We did all of this by following the precise rules of classical logic described formally in section 12.7 (Deductive systems). This reasoning emulates modus ponens.

Now we will show that every nonzero natural number has a predecessor. That is, that the function \( p \) is not defined only at \( 0 \). To highlight the logical structure of this proof, we will use first-order natural deduction rather than work with the model \( \mathbb{N} \) of PA.

Denote by \( \theta \) the formula

\[
\xi \equiv \exists \xi . (\xi \equiv s(\xi)).
\]

Clearly \( \xi \) is the only free variable in \( \theta \). We want to derive the formula \( \forall \eta . \theta[\xi \mapsto \eta] \) from the axioms of PA.

In this part of the proof we will use inductive reasoning. This will highlight that (PA3) is not an axiom schema about specifying properties, but rather about introducing a proof technique that does not hold for general logical theories. We will not attempt to prove \( \forall \eta . \theta[\xi \mapsto \eta] \) directly. Instead, we will prove a more complicated formula that is easier to prove and then by one of the many induction principles, it will follow that our desired result holds.

We can deduce the following logical theorem:

\[
\begin{align*}
\frac{(\xi \equiv s(\xi))[\xi \mapsto \eta, \xi \mapsto s(\eta)]}{(\exists^+)} \\
\frac{(\exists \xi . (\xi \equiv s(\xi))[\xi \mapsto s(\eta)])}{(\exists^+)} \\
\frac{\theta[\xi \mapsto s(\eta)]}{(\forall^+)} \\
\frac{\theta[\xi \mapsto \eta] \to \theta[\xi \mapsto s(\eta)]}{(\forall^+)} \\
\frac{\forall \eta . (\theta[\xi \mapsto \eta] \to \theta[\xi \mapsto s(\eta)])}{(\forall^+)} \\
\frac{\theta[\xi \mapsto 0] \land \forall \eta . (\theta[\xi \mapsto \eta] \to \theta[\xi \mapsto s(\eta)])}{(\forall^+)}
\end{align*}
\]

This is precisely the antecedent of the instance of (PA3) with \( \varphi = \theta \). By theorem 882 (Syntactic deduction theorem) we have

\[
(PA3) \vdash \forall \eta . \theta[\xi \mapsto \eta].
\]

10
When interpreted in \( \mathbb{N} \), this formula \( \forall \eta . \emptyset [\xi \mapsto \eta] \) simply states that every natural number is either zero or has a predecessor. The statement does not concern itself with uniqueness nor with whether 0 has a predecessor.

But we have already shown uniqueness — the predecessor function \( p \) is a partial single-valued function. And we have shown that \( p \) is not defined at zero. The last part of the proof shows \( p \) is defined for all nonzero values.

We may choose to define \( p \) at zero by giving it a sentinel value. This is precisely the technique we use in proposition 989 (b) to show that \( s \) has a left inverse if it is injective. We can use only (PA1) to show that \( s \) is injective and then pick \( p \) to be any of its left inverses. We also want \( p \) to be as close as possible to a right inverse, however. The latter, as we have seen, is more tricky.

**Proposition 6.** The natural numbers \( \mathbb{N} \) with addition form a cancellative commutative zero-sumfree monoid with 0 as the identity.

Furthermore, the sum of two natural numbers is nonzero if and only if both numbers are nonzero, that is,

\[
 n + m = 0 \text{ if and only if } n = 0 \text{ and } m = 0. \tag{1}
\]

This generalizes to proposition 1079 and proposition 1079.

**Proof.**

**Proof of commutativity.** Consider the sum \( n + m \). We use induction on \( m \) to prove its commutativity.

- If \( m = 0 \), nested induction by \( n \) yields:
  - If \( n = m = 0 \), clearly \( n + m = 0 + 0 = m + n \).
  - If the inductive hypothesis holds for its predecessor \( p(n) \),
    
    \[
    n + m = n + 0 = s(p(n)) + 0 \overset{(PA4)}{=} s(p(n)) \overset{(PA4)}{=} s(p(n) + 0) = s(p(n) + m) \overset{\text{ind}}{=} s(m + p(n)) \overset{(PA4)}{=} m + s(p(n)) = m + n.
    \]

- If \( m \neq 0 \) and if the inductive hypothesis holds for \( p(m) \), (PA1) yields that \( n + m = m + n \) if and only if \( n + p(m) = p(m) + n \). But the last equality is satisfied because of the inductive hypothesis, hence commutativity of \( n \) and \( m \) follows.

**Proof of associativity.** Fix natural numbers \( n, m, k \). We will prove associativity by induction on \( k \). If \( k = 0 \), we have

\[
 (n + m) + 0 \overset{(PA4)}{=} n + m = n + (m + 0).
\]

If \( k \neq 0 \), the proof follows directly from (PA1) as in the proof of commutativity.
Proof of identity. We have \( n + 0 = n \) by (PA4) and \( 0 + n = n \) by commutativity.

Proof of cancellation. Let \( n + k = m + k \). We will prove that \( n = m \) by induction. This is obvious for \( k = 0 \). For \( k \neq 0 \) we have

\[
  n + s(p(k)) = n + k = m + k = m + s(p(k)),
\]

which by (PA5) is equivalent to \( s(n + p(k)) = s(m + p(k)) \).

By (PA1), we have \( n + p(k) = m + p(k) \). The inductive hypothesis implies that \( n = m \).

Proof of zerosumfree. We will use induction on \( m \) in \( n + m = 0 \).

- If \( m = 0 \), then \( n + m = n \) by (PA4), and hence \( n = 0 \).
- If \( m > 0 \), then \( n + m = s(n + p(m)) \) by (PA5), and by (PA2), \( n + m \neq 0 \). Therefore, \( n + m = 0 \) if and only if \( n = m = 0 \).

Remark 7. Multiplication in commutative monoids (i.e. monoid exponentiation) is defined in definition 447 (e) for a natural number and a monoid member. It just happens that, by proposition 6, the natural numbers are themselves a monoid. We cannot rely on proposition 444, however, if we want to avoid circular definitions and proofs.

Having multiplication as part of the signature of Peano arithmetic allows us to avoid this circularity.

Proposition 8. The natural numbers \( \mathbb{N} \) with multiplication form a commutative monoid with 1 as the identity.

When combined with addition, the natural numbers become an entire commutative semiring.

This generalizes to proposition 1076 and proposition 1080.

Proof.

Proof of identity. Multiplication by 1 on the right preserves any natural number:

\[
  n \cdot 1 \overset{(PA7)}{=} n \cdot 0 + n \overset{(PA4)}{=} 0 + n = n.
\]

Multiplication from the left is handled by induction. Indeed, the case \( n = 0 \) is trivial and for nonzero \( n \) we have

\[
  1 \cdot n \overset{(PA7)}{=} 1 \cdot p(n) + 1 \overset{(PA6)}{=} p(n) + 1 = n.
\]

Proof of distributivity. We will prove that \( (n + m)k = n \cdot k + n \cdot k \) with induction on \( k \).

If \( k = 0 \),

\[
  (n + m) \cdot 0 \overset{(PA6)}{=} 0 \overset{(PA4)}{=} 0 + 0 \overset{(PA6)}{=} n \cdot 0 + m \cdot 0.
\]

For all nonzero \( k \), if the inductive hypothesis holds for \( p(k) \), then

\[
  (n + m) \cdot k = (n + m) + (n + m) \cdot p(k) \overset{\text{ind}}{=}\n
  = (n + m) + n \cdot p(k) + n \cdot p(k) =\n
  = (n + n \cdot p(k)) + (m + m \cdot p(k)) \overset{(PA7)}{=}\n
  = n \cdot k + m \cdot k.
\]
Proof of associativity. With distributivity proven, associativity of multiplication follows by induction. Indeed,

\[(n \cdot m) \cdot k = n \cdot (m \cdot k)\]

is trivially satisfied for \(k = 0\) and for all nonzero \(k\), whenever the inductive hypothesis holds for all \(n, m \in \mathbb{N}\), it follows that

\[
\begin{align*}
(n \cdot m) \cdot k & \overset{(\text{PA7})}{=} n \cdot m + (n \cdot m) \cdot p(k) \overset{\text{ind.}}{=} \\
& \overset{(\text{PA7})}{=} n \cdot m + n \cdot (m \cdot p(k)) \overset{(\text{186})}{=} \\
& = n \cdot (m + m \cdot p(k)) \overset{(\text{PA7})}{=} \\
& = n \cdot (m \cdot k).
\end{align*}
\]

Proof of commutativity. By induction on \(m\) we prove

\[n \cdot m = m \cdot n.\]

The base case is trivial for nonzero \(m\), if \(n \cdot p(m) = p(m) \cdot n\) for all \(n \in \mathbb{N}\), then

\[n \cdot m \overset{(\text{PA7})}{=} n + n \cdot p(m) \overset{\text{ind.}}{=} n + p(m) \cdot n \overset{(\text{S})}{=} (1 + p(m)) \cdot n \overset{(\text{PA4})}{=} m \cdot n.
\]

Proof of no zero divisors. We will now show that \(\mathbb{N}\) has no zero divisors.

If \(m = 0\), then \(n \cdot m = 0\) by (PA6). If \(n = 0\), then by induction on \(m\) we can easily show that \(n \cdot m = n \cdot p(m) + n = 0 + 0 = 0\).

Conversely, let \(n \cdot m = 0\). By induction on \(m\), either \(m = 0\) or we have \(n \cdot m = n \cdot p(m) + n\), in which case (1) both \(n \cdot p(m)\) and \(n\) are zero. Thus, if we assume that \(m \neq 0\), then we can conclude that \(n = 0\).

Remark 9. In [Pea89, p. 1], Peano defined an “\(n \mapsto n + 1\)” operation rather than a successor operation. It has since become common practice to instead define a “successor” operation, define addition and then show that the two are compatible:

\[n + 1 = s(n + 0) = s(n).\]

The predecessor operation then corresponds to integer subtraction by 1. We avoid subtraction in this subsection — we are only interested in the fact that every nonzero natural number \(n\) has a predecessor \(m\) such that \(n = m + 1\).

It is dangerous to conflate \(n + 1\) and \(s(n)\) until we have proved the familiar properties of addition. We have already done, so in proposition 6, however, and will further avoid mentioning directly the operations \(s(n)\) and \(p(n)\).

See remark 1069 for the more general case of ordinal addition.

Definition 10. Although this is sometimes stated as part of PA, e.g. [Aut20, exmpl. 17.6], we can define the familiar order \(\leq\) on the natural numbers via addition as the predicate formula

\[\alpha \leq \beta := \exists \xi . (\alpha + \xi = \beta).\] (2)
We use the infix notation for convenience, however we do not assume that ≤ is part of the language of PA (as explained in remark 825 (e)).

The following relation
\[ \alpha < \beta := \exists \xi \neq 0. (\alpha + \xi = \beta) \]  
(3)
is then connected to ≤ via (494).

For the specific model of PA based on the smallest inductive set \( \omega \), the latter relation < is equivalent to ∈ as discussed in remark 1002.

We will show in proposition 11 that \( \mathbb{N} \) is total ordered (even well-ordered) with ≤ as the nonstrict order and < as the strict order.

**Proof.** We will show that \( n < m \) if and only if \( n \leq m \) and \( n \neq m \).

**Proof of sufficiency.** Assume that \( n < m \). Then there exists some nonzero \( a \) such that \( n + a = m \). In particular, we have \( n \leq m \). If we suppose that \( n = m \), then \( n + a = m \) and since addition is cancellative, it would follow that \( a = 0 \), contradicting the assumption that \( a \) is nonzero.

Therefore, \( n \leq m \) and \( n \neq m \).

**Proof of necessity.** Assume that \( n \leq m \) and \( n \neq m \). Then there exists some \( a \) such that \( n + a = m \) If we suppose that \( a = 0 \), then we would obtain that \( n = m \), which would contradict our choice of \( n \) and \( m \).

Therefore, \( n < m \). \( \square \)

**Proposition 11.** The natural numbers are well-ordered by the relation < defined by (3).

Furthermore, \( \mathbb{N} \) is an ordered semiring. That is, the nonstrict order ≤ is compatible with addition and multiplication.

**Proof.** As discussed in definition 996, in order to show that < well-orders \( \mathbb{N} \), we only need to show that < is transitive, satisfies trichotomy and does not allow an infinitely descending chain.

**Proof of transitivity.** Let \( n < m \) and \( m < k \). Then there exist nonzero numbers \( a \) such that \( n + a = m \) and \( b \) such that \( m + b = k \). Thus, \( n + a + b = k \), which demonstrates that \( n \leq k \). Furthermore, because \( \mathbb{N} \) is zerosumfree, it also follows that \( a + b \neq 0 \).

Therefore, \( n < k \).

**Proof of trichotomy.** Let \( n \) and \( m \) be natural numbers.

We have already shown in definition 10 that due to (494) the equality \( n = m \) holds if and only if neither \( n < m \) nor \( n > m \).

Aiming at a contradiction, suppose that both \( n < m \) and \( n > m \) hold. There must exist nonzero numbers \( a \) and \( b \) such that \( n + a = m \) and \( n = m + b \). Then

\[ n = m + b = (n + a) + b. \]

Since addition is cancellative, we have \( a + b = 0 \). Therefore, \( n = m \), which as we have shown is incompatible with neither \( n < m \) nor \( n > m \).

Therefore, at most one of the three conditions \( n = m \), \( n < m \) or \( n > m \) holds.
We will use induction on $m$ to show that at least one of the conditions hold. If $m = 0$, then either $n = 0$ and $m = n$ or $n \neq 0$ and $m < n$. Now suppose that the inductive hypothesis holds for $m$. We will show that it also holds for $m + 1$.

- If $n = m$, then clearly $n < m + 1$.
- If $n < m$, then since $m < m + 1$ by transitivity of $<$, we have $n < m + 1$.
- If $n > m$, then there exists some nonzero $a$ such that $n = m + a$. If $a = 1$, then $n = m + 1$. If $a$ is neither 0 nor 1, then $n > m + 1$.

**Proof of well-foundedness.** We will show by induction on $n$ that an infinitely descending chain ending at $n$ cannot exist.

If $n = 0$, by (PA2) $n$ has no predecessor and thus there cannot exist a natural number $m$ such that $m < n$.

Now assume that the inductive hypothesis holds for $n$ and suppose that there exists an infinitely descending chain ending in $n + 1$:

$$\cdots < k < m < n + 1. \tag{4}$$

If $m = n$, it follows that

$$\cdots < k < n$$

is an infinitely descending chain ending in $n$.

If $m < n$, then

$$\cdots < k < m < n$$

is again an infinitely descending chain ending in $n$.

By the inductive hypothesis, a chain ending at $n$ cannot exist, therefore neither does (4).

**Proof of compatibility with addition.** We will show that the nonstrict order is compatible with addition in $\mathbb{N}$. Let $n \leq m$ and let $k$ be an arbitrary natural number. Since $n \leq m$, there exists a number $a$ such that $n + a = m$. Then

$$m + k = (n + a) + k = (n + k) + a.$$ 

Therefore,

$$n + k \leq m + k.$$ 

**Proof of compatibility with multiplication.** If $n \geq 0$ and $m \geq 0$, then $n \cdot m \geq 0$ for the simple reason that all natural numbers are greater than or equal to zero.

**Proposition 12.** The semiring $\mathbb{N}$ of natural numbers is a bounded lattice with respect to semiring divisibility. Explicitly:

(a) The join of $n$ and $m$ is their least common multiple $\text{lcm}\{n, m\}$ (defined via algorithm 689 (Euclidean algorithm) and proposition 673).

(b) The bottom element is 1 since 1 divides every natural number.
Figure 1: A spatial Hasse diagram for a fragment of the natural number divisibility lattice

\[
\begin{align*}
4 & \quad 2 \quad 6 \\
9 & \quad 3 \\
\langle 1 \rangle & = \mathbb{N} \\
\langle 2 \rangle & = \langle 4 \rangle + \langle 6 \rangle \\
\langle 3 \rangle & = \langle 6 \rangle + \langle 9 \rangle \\
\langle 4 \rangle & = \langle 2 \rangle \cap \langle 3 \rangle \\
\langle 9 \rangle & \\
\end{align*}
\]

Figure 2: A comparison of the divisibility lattice of \( \mathbb{N} \) and the lattice of ideals of \( \mathbb{Z} \).

(c) Dually, the meet of \( n \) and \( m \) is their greatest common divisor \( \gcd\{n, m\} \) (defined via algorithm 689 (Euclidean algorithm)).

(d) The top element is 0 since every natural number divides 0.

Furthermore, divisibility is compatible with the standard ordering in the sense that \( n \mid m \) implies \( n \leq m \).

Proof. By proposition 539, divisibility is a preorder.

Proof of antisymmetry. If \( n \mid m \) and \( m \mid n \), there exist numbers \( a \) and \( b \) such that \( n = ay \) and \( m = bx \). Then \( n = abx \), and we can cancel \( n \) to obtain that \( ab = 1 \). But 1 is the only unit in \( \mathbb{N} \), hence \( a = b = 1 \), and thus \( n = m \).

Proof of lattice structure. By algorithm 689 (Euclidean algorithm), every pair of integers has a positive greatest common divisor, and also a least common multiple.

By proposition 558 (c), the lattice of principal ideals in \( \mathbb{N} \) must be dual to it. Indeed, by
Theorem 684 (Bezout’s identity), we have that
\[ \langle n \rangle + \langle m \rangle = \langle \gcd(n, m) \rangle, \]
and by Proposition 558 (c), \( \langle n \rangle \cap \langle m \rangle \) contains the common multiples of \( n \) and \( m \), hence
\[ \langle n \rangle \cap \langle m \rangle = \langle \lcm(n, m) \rangle. \]

**Proof that the order are compatible.** If \( n \mid m \), then there exists a positive natural number \( a \) such that \( an = m \). We have
\[ an \overset{(PA)}{=} (a - 1)n + n = m. \]

Thus, by (2), \( n \leq m \). \( \square \)
1.2. Integers

**Definition 13.** The ring \( \mathbb{Z} \) of integers is defined as the Grothendieck completion of the commutative semiring \( \mathbb{N} \) of zero-based natural numbers.

**Lemma 14.** Consider the canonical embedding \( \iota : \mathbb{N} \to \mathbb{Z} \). For every nonzero integer \( n \), there either exists a unique natural number \( a \) such that \( n = \iota(a) \), or a unique natural number \( a \) such that \( n = -\iota(a) \).

**Proof.**

**Proof of existence.** Due to the trichotomy on natural numbers shown in definition 10, we have the following mutually exclusive possibilities:

- If \( a = b \), then \( x = [(a, a)] = [(0_N, 0_N)] \), hence \( x \) is zero.
- If \( a < b \), by definition 10, there exists some positive natural number \( c \) such that \( a + c = 0_N + b \). Then \( x = [(a, b)] = [(c, 0_N)] = \iota(c) \).
- If \( a > b \), there exists some natural number \( d \) such that \( 0_N + a = b + d \). Then \( x = [(a, b)] = [(0_N, d)] = -[(d, 0_N)] = -\iota(d) \).

**Proof of uniqueness.** If \( \iota(a) = \iota(b) \), then there exists some natural number \( u \) such that \( a + 0_N + u = 0_N + b + u \).

By proposition 6, \( \mathbb{N} \) is cancellative, so \( a = b \).

**Proof of exclusive conditions.** Finally, suppose that \( n = \iota(a) = -\iota(b) \). Then \( a + b = 0_N \). Since \( \mathbb{N} \) is zerosumfree by proposition 6, it follows that \( a = b = 0_N \), and hence \( n \) is zero. Therefore, if \( n \) is nonzero, either \( n = \iota(a) \) for some \( a \) or \( n = -\iota(b) \) for some \( b \). \( \Box \)

**Definition 15.** Consider the canonical embedding \( \iota : \mathbb{N} \to \mathbb{Z} \). Define the signum function \( \text{sgn} : \mathbb{Z} \to \mathbb{Z} \)

\[
\text{sgn}(n) := \begin{cases} 
0, & n = \iota(0_N), \\
1, & n = \iota(a) \text{ for some nonzero natural number } a, \\
-1, & n = -\iota(a) \text{ for some nonzero natural number } a.
\end{cases}
\]

This is a well-defined total function due to lemma 14.

We then classify integers based in their sign as follows

<table>
<thead>
<tr>
<th>Positive</th>
<th>( \text{sgn}(n) = 1 )</th>
<th>Nonpositive</th>
<th>( \text{sgn}(n) \neq 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>( \text{sgn}(n) = -1 )</td>
<td>Nonnegative</td>
<td>( \text{sgn}(n) \neq -1 )</td>
</tr>
<tr>
<td>Zero</td>
<td>( \text{sgn}(n) = 0 )</td>
<td>Nonzero</td>
<td>( \text{sgn}(n) \neq 0 )</td>
</tr>
</tbody>
</table>

**Proposition 16.** Integer signs has the following basic properties:
(a) The ring of integers is entire, i.e. an integral domain.

(b) If two integers have the same sign, their sum also has the same sign.

(c) $\text{sgn}(nm) = \text{sgn}(n) \cdot \text{sgn}(m)$.

(d) $n$ is zero if and only if $-n$ is zero.

(e) $n$ is positive if and only if $-n$ is negative.

Proof.

Proof of 16 (a). Suppose that $nm = 0$.

- If $n = \iota(a)$ and $m = \iota(b)$ are nonnegative, since $\mathbb{N}$ is entire and since $\iota$ is a homomorphism, then $nm = \iota(a) \cdot \iota(b) = \iota(ab) = \iota(0_{\mathbb{N}})$ implies that at least one of $a$ or $b$ is zero.
- If $n = \iota(a)$ is nonnegative and $m = -\iota(b)$ is nonpositive, this reduces to the previous case since $nm = 0 = n(-m)$.
- The other cases are similar.

Therefore, $n = 0$ or $m = 0$.

Proof of 16 (b). Fix two integers $n$ and $m$.

- If $n$ and $m$ are both zero, their sum is also zero.
- If $n = \iota(a)$ and $m = \iota(b)$ are both positive, then $n + m = \iota(a + b)$, so $n + m$ is also positive.
- If $n = -\iota(a)$ and $m = -\iota(b)$ are both negative, then $n + m = -\iota(a + b)$, so $n + m$ is also negative.

Proof of 16 (c). Fix two integers $n$ and $m$.

- If either $n$ or $m$ is zero, the product $nm$ is again zero, so
  $$\text{sgn}(nm) = \text{sgn}(n) \cdot \text{sgn}(m) = 0.$$
- If both $n = \iota(a)$ and $m = \iota(b)$ are positive, then $nm = \iota(ab)$ is nonnegative. Furthermore, $nm$ cannot be zero because $\mathbb{Z}$ is an integral domain. Hence,
  $$\text{sgn}(nm) = \text{sgn}(n) \cdot \text{sgn}(m) = 1.$$
• If \( n = \iota(a) \) is positive and \( m = -\iota(b) \) is negative, then \( n(-m) = \iota(ab) \) is positive. Furthermore, \( nm = -n(-m) = \iota(ab) \), and hence \( nm \) is negative. Then

\[
\text{sgn}(nm) = \frac{\text{sgn}(n) \cdot \text{sgn}(m)}{1} = -1.
\]

The case where \( n \) is negative and \( m \) is positive follows from this one due to commutativity.

• If both \( n = -\iota(a) \) and \( m = -\iota(b) \) are negative, then \((-n)(-m) = \iota(ab) \) is positive. Furthermore, \( nm = (-n)(-m) = -\iota(ab) \), and hence \( nm \) is positive.

\[
\text{sgn}(nm) = \frac{\text{sgn}(n) \cdot \text{sgn}(m)}{-1} = 1.
\]

**Proof of 16 (d).** This is actually a statement about group inverses because \( 0 = -0 \).

**Proof of 16 (e).** From proposition 16 (c) it follows that \( \text{sgn}(-n) = \text{sgn}(-1) \cdot \text{sgn}(n) = -\text{sgn}(n) \).

\[\square\]

**Definition 17.** We extend the natural number ordering \( \leq_{\mathbb{N}} \) to the integers \( \mathbb{Z} \) via the following truth table:

<table>
<thead>
<tr>
<th>( n )</th>
<th>( m )</th>
<th>( n \leq m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonnegative</td>
<td>nonnegative</td>
<td>( n \leq_{\mathbb{N}} m )</td>
</tr>
<tr>
<td>nonnegative</td>
<td>negative</td>
<td>( F )</td>
</tr>
<tr>
<td>negative</td>
<td>nonnegative</td>
<td>( T )</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>( -m \leq_{\mathbb{N}} -n )</td>
</tr>
</tbody>
</table>

**Proposition 18.** Integer ordering has the following basic properties:

(a) \( 0 \leq n \) if and only if \( n \) is positive.

(b) \( 0 \geq n \) if and only if \( n \) is negative.

(c) \( n \leq m \) if and only if \( -m \leq -n \).

(d) It is a total order.

Proof.

**Proof of 18 (a).** Trivial.

**Proof of 18 (b).** Trivial.

**Proof of 18 (c).** Let \( n \leq m \).

• If \( n \) and \( m \) are both either nonnegative or negative, then \( -m \leq -n \) by definition.

• If \( n \) is negative and \( m \) is not, then by proposition 16 (e) \(-m \) is negative and \(-n \) is not. Hence \(-m \leq -n \).
The converse direction of the proof is identical.

**Proof of 18 (d).** The order \( \leq \) is clearly defined for every pair of integers, so it remains to show that \( \leq \) is a partial order.

**Proof of reflexivity.** If \( n \) is positive, then \( n \leq n \) since \( n \leq \mathbb{N} \). Otherwise, \( n \leq n \) since \(-n \leq \mathbb{N} -n\).

**Proof of antisymmetry.** Suppose that \( n \leq m \) and \( m \leq n \).

- If \( n \) and \( m \) are both nonnegative, then \( n = m \) from the antisymmetry of the natural number ordering.
- If \( n \) and \( m \) are both negative, then \(-n = -m \) from the antisymmetry of the natural number ordering.
- If \( n \) is nonnegative and \( m \) is negative, we have \( m \leq n \) but not \( n \leq m \). This contradicts our assumption.
- If \( n \) is negative and \( m \) is nonnegative, we have \( n \leq m \) but not \( m \leq n \). This contradicts our assumption.

**Proof of transitivity.** Suppose that \( n \leq m \) and \( m \leq k \).

- If \( n, m \) and \( k \) are nonnegative, then \( n \leq k \) from the transitivity of the natural number ordering.
- If \( n, m \) and \( k \) are negative, then \(-k \leq -n \) and, by proposition 18 (c), \( n \leq k \).
- If \( n \) is negative and \( k \) is nonnegative, then \( n \leq k \) by definition.

\[ \square \]

**Definition 19.** We define the function

\[
|\cdot| : \mathbb{Z} \to \mathbb{N}
\]

\[
|n| := \begin{cases} 
  n, & n \geq 0 \\
  -n, & n < 0. 
\end{cases}
\]

We call \( |\cdot| \) the **absolute value function** in \( \mathbb{Z} \). It does satisfy the absolute value axioms from definition 156. A proof would be cyclic, however, since the real numbers rely on an extension of \(|\cdot|\).

**Algorithm 20** (Integer division). Fix two integers \( n \) and \( m \), and assume that \( m \) is nonzero. Define

\[
q := \text{sgn}(nm) \cdot \max\{y \geq 0 : y|m| < |n|\}
\]

\[
r := n - mq
\]

Then \( q \) and \( r \) and the unique integers such that \( |r| < |m| \) and

\[
n = mq + r.
\]

We will use the notation \( \text{quot}(n, m) \) and \( \text{rem}(n, m) \) from definition 687.
Proof.

**Proof of correctness.** The case $n = 0$ is trivial so assume that $n \neq 0$. We have

<table>
<thead>
<tr>
<th>$n$</th>
<th>$m$</th>
<th>$q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>positive</td>
<td>$\max{y \geq 0 : (n - m \leq y) + m &lt; +n}$</td>
</tr>
<tr>
<td>positive</td>
<td>negative</td>
<td>$-\max{y \geq 0 : (n + m \leq y) - m &lt; +n}$</td>
</tr>
<tr>
<td>negative</td>
<td>positive</td>
<td>$-\max{y \geq 0 : (m - n \leq y) + m &lt; -n}$</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
<td>$\max{y \geq 0 : (n - m \leq y) - m &lt; -n}$</td>
</tr>
</tbody>
</table>

It follows that $|n - mq| < |m|$.

**Proof of uniqueness.** Suppose that $n = mq + r = mq' + r'$, where $|r| < |m|$ and $|r'| < |m|$. Then

$$m(q - q') = -(r - r').$$

Thus, $m$ divides $r - r'$. Then $m$ divides $r$ and $r'$ contradicting the assumption that $|r| < |m|$.

**Proposition 21.** The ring of integers is an Euclidean domain with division given by algorithm 20 (Integer division) and degree function $\cdot$.

**Proof.** By proposition 16 (a), $\mathbb{Z}$ is an integral domain. The Euclidean domain structure is described by algorithm 20 (Integer division).

**Remark 22.**

- By proposition 21, $\mathbb{Z}$ is a Euclidean domain.

- By proposition 688 (a), $\mathbb{Z}$ is a principal ideal domain.

- By proposition 683 (b), $\mathbb{Z}$ is a unique factorization domain.

- By proposition 681 (b), $\mathbb{Z}$ is a GCD domain.

**Remark 23.** We discuss in remark 668 and remark 669 how, in general GCD domain, the greatest common divisors are not unique. The greatest common divisor of is, by convention, positive. This leaves a canonical choice for both the greatest common divisor and the least common multiple. By proposition 12, the divisibility order of positive integers is compatible with the usual integer ordering.

Algorithm 689 (Euclidean algorithm) allows us to explicitly compute both GCDs and LCMs.

**Definition 24.** A prime number is an integer greater than 1 whose only proper divisor is 1. Non-prime integers greater than 1 are called composite numbers.

**Remark 25.** The definition of a prime number given in definition 24 is standard, however it seems quite inconsistent with section 10.9 (Integral domains).

First, section 10.9 (Integral domains) actually defines irreducible elements rather than prime elements of the domain $\mathbb{Z}$. Second, if $p$ is a prime number, $-p$ is also a prime number.
Fortunately, prime and irreducible elements coincide in GCD domains by proposition 674 (c) and proposition 674 (c). Unfortunately, calling negative prime elements of \( \mathbb{Z} \) "prime numbers" is not accepted.

Coprime integers are, fortunately, defined as in general GCD domains via definition 676.

**Lemma 26** (Euclid’s lemma). If \( p \) is a prime number, then \( p \mid nm \) implies \( p \mid n \) or \( p \mid m \).

**Proof.** Since \( \mathbb{Z} \) is a GCD domain, the lemma follows from proposition 674 (c). \( \square \)

**Theorem 27** (Fundamental theorem of arithmetic). Every integer greater than 1 can be factored into a product of prime powers.

**Proof.** We have discussed in remark 22 that \( \mathbb{Z} \) is a unique factorization domain. \( \square \)

**Definition 28.** For any positive integer \( n \), denote by \( \varphi(n) \) the number of strictly smaller than \( n \) positive integers that are coprime to \( n \). We call \( \varphi : \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0} \) Euler’s totient function.

**Proposition 29.** Euler’s totient function \( \varphi \) has the following basic properties:

(a) \( \varphi(1) = 0 \).

(b) If \( p \) is prime, then \( \varphi(p) = p - 1 \).

(c) The multiplicative group \( \mathbb{Z}_n^\times \) of the ring \( \mathbb{Z}_n \) of integers modulo \( n > 1 \) has order \( \varphi(n) \).

**Proof.**

**Proof of 29 (a).** There are no positive integers smaller than 1.

**Proof of 29 (b).** Every positive integer smaller than \( p \) is coprime to \( p \), and there are exactly \( p - 1 \) positive integers smaller than \( p - 1, 2, \ldots, p - 1 \).

**Proof of 29 (c).** Follows from corollary 685. \( \square \)

**Theorem 30** (Euler’s totient theorem). For positive coprime integers \( n \) and \( x \), we have

\[
x^{\varphi(n)} \equiv 1 \pmod{n},
\]

where \( \varphi \) is Euler’s totient function.

**Proof.** This is vacuous for \( n = 1 \) since all integers are equal modulo 1.

Suppose that \( n > 1 \). First, use algorithm 20 (Integer division) to obtain integers \( q \) and \( y < n \) such that

\[
x = nq + y.
\]

Since \( x \) is, by assumption, coprime with \( n \), then \( y \) is also coprime with \( n \). Indeed, every common divisor \( d \) of \( y \) and \( n \) is also a common divisor \( x \), and the largest such possible value is \( \gcd(n, x) = 1 \).

Now consider the multiplicative group \( \mathbb{Z}_n^\times \) of the ring \( \mathbb{Z}_n \) of integers modulo \( n \) and the cyclic subgroup \( \{1, y, y^2, \ldots\} \) (modulo \( n \)). It is necessarily finite as a subgroup of \( \mathbb{Z}_n^\times \). Furthermore,
by theorem 468 (Lagrange’s theorem for groups), its order $k$ divides the order of $\mathbb{Z}_n^\times$. By proposition 29 (c), the order of $\mathbb{Z}_n^\times$ is $\varphi(n)$.

We have $y^k \equiv 1 \pmod{n}$ since $k$ is the order of a cyclic group. If $\varphi(n) = km$, then

$$y^{\varphi(n)} = y^{km} \equiv (y^k)^m \equiv 1^m \pmod{n}.$$ 

\[\square\]

**Corollary 31.** Given positive integers $n$ and $m$, we can apply algorithm 20 (Integer division) to obtain $n = q\varphi(m) + r$, where $\varphi$ is Euler’s totient function.

Then, for a positive integer $x$ coprime to $m$, we have

$$x^n \equiv x^r \pmod{m},$$

*Proof.* By theorem 30 (Euler’s totient theorem), $x^{\varphi(m)} \equiv 1 \pmod{m}$. Then

$$x^n = (x^{\varphi(m)})^q x^r \equiv x^r \pmod{m}.$$ 

\[\square\]

**Example 32.** The integers 9 and 10 are coprime. We have $\varphi(9) = 6$ and $1000 = 166 \cdot 6 + 4$.

By corollary 31,

$$9^{1000} \equiv 9^4 \equiv 6561 \equiv 1 \pmod{10}.$$ 

We can thus vastly simplify finding the last digit of the decimal representation of $9^{1000}$.

**Theorem 33** (Fermat’s little theorem). For a prime number $p$ and for any positive integer $x$, we have

$$x^p \equiv x \pmod{p}.$$ 

*Proof.* If $p \mid x$, then both $x^p$ and $x$ and congruent to 0 modulo $p$.

Otherwise, by theorem 30 (Euler’s totient theorem), we have $x^{\varphi(p)+1} \equiv x \pmod{p}$, and by proposition 29 (b), we have $\varphi(p) + 1 = p$. 

\[\square\]
1.3. Real numbers

**Definition 34.** The rational numbers $\mathbb{Q}$ are the field of fractions of the integers. Both operations from $\mathbb{Z}$ are inherited in $\mathbb{Q}$ and all nonzero elements in $\mathbb{Q}$ are now invertible, which makes $\mathbb{Q}$ a field.

**Definition 35.** The real numbers $\mathbb{R}$ are the metric space completion of $\mathbb{Q}$ with respect to the absolute value. Unfortunately, real numbers are used for defining metric spaces, so we cannot rely on the theory of metric spaces. This can be circumvented by

1. Regarding $\mathbb{Q}$ as a uniform space.
2. Using uniform space completion to obtain $\mathbb{R}$.
3. Defining metric spaces.
4. Showing that $\mathbb{R}$ is a metric space.
5. Using definition 36 (b) to automatically verify that $\mathbb{R}$ is complete as a metric space.

**Definition 36.** We are sometimes interested in extended real numbers. These can be any of the three sets

- $\mathbb{R} \cup \{+\infty\}$,
- $\mathbb{R} \cup \{-\infty\}$,
- $\mathbb{R} \cup \{-\infty, +\infty\}$,

where $-\infty$ and $+\infty$ are both sentinel values that act as the greatest and/or least real number. We generally avoid performing arithmetic operations on $\pm \infty$, however it is sometimes convenient to define

$$x + (+\infty) := +\infty, \ x \in \mathbb{R} \quad x \cdot (+\infty) := +\infty, \ x \in \mathbb{R}$$

We leave the operations

$$(-\infty) + (+\infty) \quad (-\infty) \cdot (+\infty)$$

undefined.

With these operations, the extended real numbers are no longer a field.

**Definition 37.** Let $x \in \mathbb{R}$ be a real number. In analogy with definition 534, we define its **floor**

$$\text{floor}(x) := \max\{n \in \mathbb{Z} : n \leq x\},$$

its **ceiling**

$$\text{ceil}(x) := \min\{n \in \mathbb{Z} : n \geq x\}$$

and its **fractional part**

$$\text{frac}(x) := x - \text{floor}(x).$$
Proposition 38. The field $\mathbb{R}$ is not algebraically closed.
In particular, the polynomial $x^2 + 1$ has no root.

Proof. Assume that $\mathbb{R}$ is algebraically closed and that the polynomial $x^2 + 1$ has at least one root. Denote one of them by $u$.

By the trichotomy of the order $<$ of $\mathbb{R}$, we have either $u < 1$ or $u > 1$ since $u \neq 1$.

If $u < 0$, then $u^2 = -1 < 0$, which is impossible because the image of $x \mapsto x^2$ is the interval $[0, \infty)$.

If $u > 0$, then $u^2 = -1 < 0 = 0$, which is also impossible because $x \mapsto x^2$ is monotone on $[0, \infty)$.

Thus, $u$ is not a root of $x^2 + 1$ and $\mathbb{R}$ is not algebraically closed. \qed

Definition 39. We define the signum function $\text{sgn} : \mathbb{R} \to \{-1, 0, 1\}$ as

$$
\text{sgn}(x) := \begin{cases} 
1, & x > 0, \\
0, & x = 0, \\
-1, & x < 0.
\end{cases}
$$
1.4. Complex numbers

**Definition 40.** We give a few equivalent definition of the field $\mathbb{C}$ complex numbers. Informally, there are numbers of the form $a + bi$, where $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$. In order to find the multiplicative inverse of the nonzero polynomial $a + bi$, we assume that division is well-defined and proceed as follows:

$$\frac{1}{a + bi} = \frac{a - bi}{(a + bi)(a - bi)} = \frac{a - bi}{a^2 + b^2}. \quad (5)$$

The closest to this informal definition is definition 40(a).

(a) The most "algebraic" way to define complex numbers is as the polynomial quotient ring

$$\mathbb{C} := \mathbb{R}[X]/(X^2 + 1).$$

Elements of $\mathbb{C}$ can be identified with real polynomials of the form $bX + a$. See ?? ([UNDEFINED]) for a broader discussion. Define $i := X$. We have

$$i \cdot i = X^2 = -1 \pmod{X^2 + 1}.$$ 

Thus, $i$ is indeed the square root of $-1$. We will write

$$a + bi = bX + a.$$ 

It is shown in ?? ([UNDEFINED]) that multiplication modulo $X^2 + 1$ gives

$$(bX + a)(dX + c) = (ad + bc)X + (ac - bd) \pmod{X^2 + 1}. \quad (6)$$

The multiplicative inverse of $a + bi$ is then indeed eq. (5).

The canonical embedding $\iota : \mathbb{R} \to \mathbb{C}$ is then the standard polynomial embedding.

(b) The complex numbers can also be defined as the matrix ring

$$\mathbb{C} := \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} : a, b \in \mathbb{R} \right\}$$

with the usual matrix multiplication. The canonical embedding $\iota : \mathbb{R} \to \mathbb{C}$ is then

$$\iota(a) := \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$$

(c) Finally, we can define $\mathbb{C}$ is the algebra obtained from the vector space $\mathbb{R}^2$ with the multiplication operation emulating eq. (6) as

$$\cdot : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$$

$$(a, b) \cdot (c, d) := (ac - bd, ad + bc).$$

The canonical embedding $\iota : \mathbb{R} \to \mathbb{C}$ is then

$$\iota(a) := (a, b).$$
We define the unary complex conjugation operation as \( a + bi \) := \( a - bi \) and the absolute value as
\[
|a + bi| := \sqrt{a^2 + b^2}.
\]

For a complex number \( z = a + bi \) we denote
\[
\text{real } z = a \quad \quad \text{imag } z = b
\]
and call them the real and imaginary parts of \( z \).

**Theorem 41** (Fundamental theorem of algebra). The field \( \mathbb{C} \) of complex numbers is algebraically closed.

**Theorem 42.** Let \( X \) be a vector space over \( \mathbb{C} \). There is a bijection between the real-valued and the complex-valued linear functionals on \( X \).

**Proof.** Let \( c : X \to \mathbb{C} \) be a complex-valued linear functional. Denote \( a(x) := \text{real } c(x) \) and \( b(x) := \text{imag } c(x) \). Then \( a : X \to \mathbb{R} \) and \( b : X \to \mathbb{R} \) are linear functionals. We will show that \( a(x) \) uniquely determines \( b(x) \) and hence \( c(x) \).

Note that \( c(ix) = a(ix) + ib(ix) = ia(x) - b(x) \). Therefore, \( b(x) = a(ix) - c(ix) \) and
\[
c(x) = a(x) + i(a(ix) - c(ix)) = a(x) - a(x) + c(x) = c(x).
\]

\( \square \)

**Remark 43.** Theorem 42 allows us to identify the dual space \( X^* \) of a complex vector space \( X \) with \( \text{hom}(X, \mathbb{R}) \) in the case of an algebraic dual or with the corresponding subspace in the case of a continuous dual space.

This allows us to reuse some of the theory for real vector spaces, for example hyperplane separation.
2. Real analysis

Real analysis is concerned with functions with values in Euclidean spaces. These include the real line $\mathbb{R}$, plane $\mathbb{R}^2$ or space $\mathbb{R}^3$, which are the classic spaces from section 8.3 (Analytic geometry in the plane).

Important aspects of real analysis include:

- **Continuity**, which we discuss extensively in section 6 (General topology) and section 7 (Metric spaces), and aggregate in section 2.1 (Topology of Euclidean spaces) and section 2.3 (Real convergence).

- **Differentiability**, which we discuss in section 4.10 (Differentiability) from section 4 (Functional analysis) and discuss briefly in section 4.10 (Differentiability), and, in a generalized form, in section 2.9 (Nonsmooth derivatives).

- Integration, which we discuss in section 2.6 (Riemann integration) and section 2.7 (Line integrals).

- Special functions, which we delegate to section 3 (Complex analysis) via section 3.7 (Special functions).

- Convex functions, which we discuss in section 2.10 (Convex functions).
2.1. Topology of Euclidean spaces

**Definition 44.** TODO: Define euclidean spaces.

**Theorem 45.** For the real numbers, the metric and order topologies coincide.

*Proof.* The metric topology $\mathcal{T}_M$ is generated by the base

$$B := \{B(x, r) : x \in \mathbb{R}, r \in \mathbb{R}_{>0}\}$$

and the order topology $\mathcal{T}_O$ is generated by the subbase

$$\mathcal{P} := \{(a, \infty) : a \in \mathbb{R}\} \cup \{(-\infty, b) : b \in \mathbb{R}\}.$$  

The inclusion $B \subseteq FI(\mathcal{P})$ is obvious since any ball $B(x, r)$ is the intersection of the two rays

$$B(x, r) = (x - r, \infty) \cap (-\infty, x + r).$$

Thus, $\mathcal{T}_M \subseteq T_O$. We now only need to show that $B$ is a base for $\mathcal{T}_O$.

Let $U \in T_O$. Since $FI(\mathcal{P})$ is a base for $\mathcal{T}_O$, there exists a family $\{U_i\}_{i \in I} \subseteq FI(\mathcal{P})$ such that

$$U = \bigcup_{i \in I} U_i.$$  

We only need to express every $U_i$ as a union of balls from $B$. There are several possibilities:

- if $U_i$ is the open interval $(a, \infty)$,
  $$(a, \infty) = \bigcup_{i=1}^{\infty} B(a + i, 1).$$

- if $U_i$ is the open interval $(-\infty, b)$,
  $$(-\infty, b) = \bigcup_{i=1}^{\infty} B(b - i, 1).$$

- if $U_i$ is the intersection $(a, \infty) \cap (-\infty, b)$, $a < b$,
  $$(a, \infty) \cap (-\infty, b) = B\left(\frac{a + b}{2}, \frac{b - a}{2}\right)$$

- if $U_i$ is the empty set,
  $$\emptyset = \bigcup \emptyset \text{ (see definition 936 (a)).}$$

Thus, $U_i$ is the union of an at most countable amount of balls. The countable union of countable sets is again countable, hence by definition 247 (a), $B$ is a base for $\mathcal{T}_O$.  

**Proposition 46.** A set in $\mathbb{R}^n$ is totally bounded if and only if it is bounded.
Proof.

**Proof of sufficiency.** Follows from proposition 379.

**Proof of necessity.** Let $A$ be a bounded set in $\mathbb{R}^n$ and let $B(x, r)$ be a ball containing $A$. Fix $\varepsilon > 0$.

Denote by $e_1, \ldots, e_n$ the basis of $\mathbb{R}^n$. Denote by $m$ the smallest integer such that $m\varepsilon \geq r$.

We can create a grid around $B(x, r)$ as follows:

Define the set

$$\left\{ x + \sum_{i=1}^{n} \left\lceil k_i\varepsilon \right\rceil e_i : \forall i = 1, \ldots, n : k_i = 1, \ldots, m \right\}.$$  

is finite. Furthermore, it is an $\varepsilon$-net of $A$. Indeed, let $y \in A$. Denote its coordinates along $e_1, \ldots, e_n$ by $y_1, \ldots, y_n$. Then $y$ is contained in the ball

$$B\left( x + \sum_{i=1}^{n} \left\lceil \text{ceil}(y_i)\varepsilon \right\rceil e_i, \varepsilon \right).$$


**Theorem 47** (Heine-Borel theorem). A set in $\mathbb{R}^n$ is compact in the sense of definition 318 if and only if it is closed and bounded.

**Proof.** Follows from proposition 46 and corollary 383 (c).

**Proposition 48.** The supremum (resp. infimum) of a set $A \subseteq \mathbb{R}$, if it exists, is equal to the supremum (resp. infimum) of $\text{cl} A$.

**Proof.**

**Proof of sufficiency.** Denote by $M$ the supremum of $A$. Assume that it is not a supremum of $\text{cl} A$, that is, there exists an upper bound $M'$ of $\text{cl} A$ such that $M' < M$. But this is impossible because $A \subseteq \text{cl} A$.

Therefore, $M$ is a supremum of $\text{cl} A$.

**Proof of necessity.** Denote by $M$ the supremum of $\text{cl} A$. Assume that it is not a supremum of $A$, that is, there exists an upper bound $M'$ of $A$ such that $M' < M$.

Let $\{x_i\}_{i=1}^{\infty} \subseteq A$ be a sequence that converges to $M$. Then

$$x_i < M' < M.$$  

By lemma 55 (a), we have $M' = M$, which contradicts our choice of $M'$. Thus, $M$ is the supremum of $A$.

**Proposition 49.** Every nonempty bounded set in $\mathbb{R}$ has a supremum and infimum.

**Proof.** Let $A \subseteq \mathbb{R}$ be a nonempty bounded set. By theorem 47 (Heine-Borel theorem), the set $\text{cl} A$ is compact. By theorem 322 (Weierstrass' extreme value theorem), the identity function $id : \mathbb{R} \rightarrow \mathbb{R}$ attains its minimum $m$ and maximum $M$ on $\text{cl} A$. Note that both $m$ and $M$ do not have to belong to $A$, but $m$ is a lower bound and $M$ is an upper bound of the set $A$. 

31
If we take any other upper bound $M'$ of $A$, then by proposition 48,

$$M' \geq \sup A = \sup \text{cl } A = M.$$

Hence, $M$ is the least upper bound of $A$.

We can analogously prove that $m$ is the greatest lower bound of $A$.  

\qed
2.2. Real-valued functions

**Definition 50.** Let \( \mathcal{F} \) be a family of functions from a set \( S \) to a ring \( R \). We say that \( \mathcal{F} \) **vanishes nowhere** if for every \( x \in S \) there exists a function \( f \in \mathcal{F} \) such that \( f(x) \neq 0 \).

**Definition 51.** Let \( X \) be an arbitrary set. The **epigraph** of the function \( f : X \to \mathbb{R} \) is defined as

\[
epi f := \{(x, r) \in X \times \mathbb{R} : r \geq f(x)\},
\]
2.3. Real convergence

**Theorem 52** (Bolzano-Weierstrass). *Every bounded sequence in \( \mathbb{R} \) has a convergent subsequence.*

**Proof.** Let \( \{x_k\}_{k=1}^{\infty} \) be a bounded sequence in \( \mathbb{R} \) and let \( a \leq b \) be a lower and upper bound, respectively. Construct the sequence \( \{F_k\}_{k=1}^{\infty} \) of closed intervals as follows: define \( \alpha_1 := a \) and \( \beta_1 := b \) and, at step \( k = 1, 2, \ldots \), put

\[
F_k := \begin{cases} 
[\alpha_k, \frac{\alpha_k + \beta_k}{2}], & \text{if } \frac{\alpha_k + \beta_k}{2} \text{ contains infinitely many sequence members,} \\
[\frac{\alpha_k + \beta_k}{2}, \beta_k], & \text{otherwise.}
\end{cases}
\]

Then put \( \alpha_{k+1} \) and \( \beta_{k+1} \) to be the endpoints of the interval \( F_k \) and repeat with \( k + 1 \) instead of \( k \). Note that for any \( k = 1, 2, \ldots \), \( \text{diam}(F_k) = \frac{1}{2} \text{diam}(F_{k-1}) \), thus \( \text{diam}(F_k) \to 0 \) as \( k \to \infty \). As in **Theorem 373** (Cantor’s nested compact theorem), it follows that if we choose a sequence \( x_k \in F_k, k = 1, 2, \ldots \), it will be a fundamental sequence. Since the space is complete, this fundamental sequence necessarily converges. \( \square \)

**Theorem 53.** *The metric space \( \mathbb{R} \) is complete.*

**Proof.** Let \( \{x_k\}_{k=1}^{\infty} \) be a fundamental sequence of real numbers. By **Proposition 369**, the sequence is bounded. By **Theorem 52** (Bolzano-Weierstrass), it has a convergent subsequence \( \{x_{k_m}\}_{m=1}^{\infty} \to x \).

By **Proposition 370**, the sequence itself has the same limit \( \lim_{k \to \infty} x_k = x \). \( \square \)

**Proposition 54.** *Fix two convergent sequences \( \{x_k\}_{k=1}^{\infty} \) and \( \{y_k\}_{k=1}^{\infty} \) of real numbers. If \( x_k \leq y_k \) for all \( k = 1, 2, \ldots \), then*

\[
\lim_{k \to \infty} x_k \leq \lim_{k \to \infty} y_k.
\]

**Proof.** Denote the respective limits by \( x \) and \( y \).

Fix \( \varepsilon > 0 \). Then by **Definition 274 (b)**, there exist indices \( k_0 \) and \( m_0 \) such that

\[
|x - x_k| < \frac{\varepsilon}{2} \quad \forall k \geq k_0 \\
|y - y_m| < \frac{\varepsilon}{2} \quad \forall m \geq m_0
\]

Take \( k \geq \max\{k_0, m_0\} \). Then \( y_k \geq x_k \) and

\[
y - x = (y - y_k) + (y_k - x_k) + (x_k - x) \geq \geq (y - y_k) + (x - x_k) >
\]

34
\[ > -\frac{\varepsilon}{2} - \frac{\varepsilon}{2} = -\varepsilon. \]

Since \( \varepsilon \) was chosen arbitrary, \( y - x \) cannot equal any negative number, because otherwise we could choose another \( \varepsilon \) smaller than the magnitude of the negative number and obtain a contradiction.

Thus, \( y \geq x \). \( \square \)

**Lemma 55** (Squeeze lemma). *Let \( I \) be a closed interval in \( \mathbb{R} \).*

(a) Let \( \{x_k\}_{k=1}^{\infty}, \{x_k^-\}_{k=1}^{\infty}, \{x_k^+\}_{k=1}^{\infty} \) be three sequences in \( I \). If both \( \{x_k^-\}_{k=1}^{\infty} \) and \( \{x_k^+\}_{k=1}^{\infty} \) converge to the same value \( \bar{x} \in I \) and if the following inequalities

\[ x_k^- \leq x_k \leq x_k^+ \]

hold for all \( k = 1, 2, \ldots \), then the “squeezed in” sequence \( \{x_k\}_{k=1}^{\infty} \) also converges to \( \bar{x} \).

(b) Let \( f, f_-, f_+ : I \to \mathbb{R} \) be three functions and let \( \bar{x} \in I \). If both limits \( \lim_{x \to \bar{x}} f_-(x) \) and \( \lim_{x \to \bar{x}} f_+(x) \) converge to the same value \( \bar{y} \in \mathbb{R} \) and if the following inequalities

\[ f_-(x) \leq f(x) \leq f_+(x) \]

hold for all \( x \in I \), then the “squeezed in” function \( f \) also converges to \( \bar{y} \) at \( \bar{x} \).

**Proof.**

**Proof of 55 (a).** Fix \( \varepsilon > 0 \). Then by definition 274 (b), there exist indices \( k^- \) and \( k^+ \) such that

\[ |\bar{x} - x_k^-| < \frac{\varepsilon}{2} \quad \forall k \geq k^- \]
\[ |\bar{x} - x_k^+| < \frac{\varepsilon}{2} \quad \forall k \geq k^+ \]

By taking \( k \geq \max\{k^-, k^+\} \), we obtain

\[ |x_k^+ - x_k^-| \leq |x_k^+ - \bar{x}| + |\bar{x} - x_k^-| < \frac{2\varepsilon}{3}. \]

Since \( |x_k^+ - x_k^-| \leq |x_k^+ - x_k^-| \), it follows that \( |x_k^+ - x_k^-| < \frac{2\varepsilon}{3} \).

Thus,

\[ |\bar{x} - x_k| \leq |\bar{x} - x_k^+| + |x_k^+ - x_k| < \varepsilon. \]

Definition 274 (b) is satisfied, hence \( \{x_k\} \) converges to \( \bar{x} \).

**Proof of 55 (b).** The proof is analogous to that of lemma 55 (a), but the machinery is different. Fix \( \varepsilon > 0 \). Then by definition 288 (a), there exist radii \( \delta^- \) and \( \delta^+ \) such that

\[ f_-(I \cap B(\bar{x}, \delta^-)) \subseteq B(\bar{y}, \frac{\varepsilon}{3}) \]
\[ f_+(I \cap B(\bar{x}, \delta^+)) \subseteq B(\bar{y}, \frac{\varepsilon}{3}) \]
Take $\delta < \min\{\delta^-, \delta^+\}$ and $x \in I \cap B(\bar{x}, \delta)$. Analogously to the proof of lemma 55 (a), we obtain the inequality

$$|f(x) - \bar{x}| \leq |f(x) - f^-(x)| + |f^-(x) - \bar{x}| < \varepsilon.$$

We conclude that

$$f(I \cap B(\bar{x}, \delta)) \subseteq B(\bar{y}, \varepsilon)$$

holds and thus by definition 288 (a), the function $f$ converges to $\bar{y}$ at $\bar{x}$. \hfill \Box

**Proposition 56.** A monotone sequence of real numbers converges if and only if it is bounded.

**Proof.**

**Proof of sufficiency.** Let $\{x_k\}_{k=1}^\infty$ be a convergent monotone sequence. Denote its limit by $x$. Fix $\varepsilon > 0$. By definition 274 (b), there exists $k_0$ such that

$$|x - x_k| < \varepsilon \quad \forall k \geq k_0.$$

Thus, $\{x_k : k \geq k_0\} \subseteq B(x, \varepsilon)$.

Also note that

$$\{x_k : k < k_0\} \subseteq B(x, \max_{i<k_0} \{|x - x_k|\}).$$

We obtained that the entire sequence

$$\{x_k : k \geq 1\} = \{x_k : k < k_0\} \cup \{x_k : k \geq k_0\}$$

is contained in a union of two balls and is therefore bounded.

**Proof of necessity.** Now let $\{x_k\}_{k=1}^\infty$ be a bounded monotone sequence. Denote its supremum by $\bar{x}$. Note that

$$|x_n - x_m| = x_n - x_m \leq \bar{x} \quad \forall n \geq m.$$

Fix $\varepsilon > 0$. Then there exists at least one element $x_{m_0} > \bar{x} - \varepsilon$ because otherwise $\bar{x}$ would not be a supremum.

Then for any index $n \geq m_0$ we have

$$|x_n - x_{m_0}| = x_n - x_{m_0} < \bar{x} - (\bar{x} - \varepsilon) = \varepsilon.$$

Thus, definition 274 (b) is satisfied and the sequence $\{x_k\}_{k=1}^\infty$ converges. \hfill \Box
Figure 3: Plot of the third partial sum of the Weierstrass function with $a = 0.9$ and $b = 7$ from $-\pi/8$ to $\pi/8$.

### 2.4. Real differentiability

**Proposition 57.** Let $U \subseteq \mathbb{R}^n$ be an open set. A real-valued function $f : U \to \mathbb{R}$ is differentiable at $x$ in the direction $h$ if and only if $\varphi(t) = f(x + th)$ is right-differentiable at 0.

**Proof.**

$$ f'(x)(h) := \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t} = \varphi'(0)(1). $$

**Example 58** (Weierstrass’ nowhere differentiable function). Let $a \in (0,1)$ and $b$ is a positive odd integer such that

$$ ab > 1 + \frac{3}{2}\pi. $$

Define the function

$$ f(x) := \sum_{k=0}^{\infty} a^k \cos(b^k \pi x). $$

Since $\cos$ is bounded for real arguments and $a \in (0,1)$, each term is uniformly bounded by 1 and by corollary 118 (Weierstrass’ series criterion), $f$ is continuous. However, it is not differentiable at any point. The proof of the latter is involved and will not be given here.

**Theorem 59.** TODO: Prove Leibniz’ rule.
2.5. Real series

**Proposition 60.** If only finitely many coefficients in a real convergent series are negative, then the series converges absolutely.

*Proof.* Let $N$ be the index of the last negative coefficient in (29). Then the series

$$\sum_{k=N+1}^{\infty} a_k$$

is absolutely convergent since every coefficient is positive. Then

$$\sum_{k=0}^{\infty} |a_k| = \sum_{k=0}^{N} |a_k| + \sum_{k=N+1}^{\infty} |a_k|$$

is convergent since the first term on the right side is a finite sum and the second is a convergent series. Hence, the series (29) converges absolutely. 

**Corollary 61.** If only finitely many coefficients in a real convergent series are positive, then the series converges absolutely.

**Theorem 62** (Riemann’s series permutation theorem). *If the real series*

$$\sum_{k=0}^{\infty} a_k$$

*is convergent, but not absolutely convergent, then for any extended real number $x \in \mathbb{R} \cup \{-\infty, +\infty\}$ there exists a permutation $p$ of the coefficients $a_0, a_1, a_2$ such that*

$$\sum_{k=0}^{\infty} p(a_k) = x.$$

*Proof.* If the series is not absolutely convergent, then there exist both infinitely many positive and infinitely many negative coefficients.

First, assume that $x$ is finite.

Define the permuted series

$$\sum_{k=0}^{\infty} b_k$$

as follows:

(a) Assign to $b_n$ only nonnegative elements of the sequence $\{a_k\}_{k=0}^{\infty}$ until $\sum_{k=0}^{n} b_k \geq x$. Then go to 62 (b).

(b) Assign to $b_n$ only negative elements of the sequence $\{a_k\}_{k=0}^{\infty}$ until $\sum_{k=0}^{n} b_k \geq x$. Then go to 62 (a).
This mutual recursion builds a series that converges to $x$ because the coefficients $\{a_k\}_{k=0}^\infty$ get arbitrarily close to each other. If $x = +\infty$, we can add positive coefficients until $\sum_{k=0}^n b_k \geq 1$, then add a single negative coefficient, then continue adding positive coefficients until $\sum_{k=0}^n b_k \geq 2$ and, so on.

If $x = -\infty$, we use the same process, but with milestones of $-1, -2, -3, \ldots$. □

**Example 63.** [Fic68b, №247] Consider the alternating harmonic series (41). Denote its sum by $\alpha$.

We can rearrange this series by repeating two negative terms and a single positive term as follows:

$$1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \cdots = \sum_{m=1}^\infty \left( \frac{1}{2m-1} - \frac{1}{4m-2} - \frac{1}{4m} \right).$$

(7)

Note that eq. (7) is equivalent to

$$\sum_{m=1}^\infty \left( \frac{1}{2m-1} - \frac{1}{4m-2} - \frac{1}{4m} \right) = \sum_{m=1}^\infty \left( \frac{1}{4m-2} - \frac{1}{4m} \right) = \frac{1}{2} \sum_{m=1}^\infty \left( \frac{1}{2m-1} - \frac{1}{42} \right) = \alpha / 2$$

**Proposition 64.** Fix two nonnegative series

$$\sum_{k=0}^\infty a_k$$  

(8)

and

$$\sum_{k=0}^\infty b_k$$  

(9)

that is, series with nonnegative real coefficients. Assume that there exists an index $K$ such that $a_k \leq b_k \ \forall k \geq K$.

We say that the series eq. (9) dominates the series eq. (8).

Then

(a) If eq. (9) converges, so does eq. (8).

(b) If eq. (8) diverges, so does eq. (9).

Proof.

**Proof of 64 (a).** Suppose that eq. (9) converges. Then by proposition 56, the sequence of partial sums is bounded. Therefore, the sequence of partial sums of eq. (8) is also bounded and, by proposition 56 again, the series is convergent.

**Proof of 64 (b).** Analogous to proposition 64 (a), but using the negation of proposition 56.
Proposition 65 (Cauchy’s root test). Consider the nonnegative series eq. (8). Put
\[ q := \limsup_{k \to \infty} \sqrt[k]{a_k}, \]
where \( q = \infty \) if the limit does not exist. Then
- If \( q < 1 \), the series converges.
- If \( q > 1 \), the series diverges.
- If the limit does not exist (e.g. if \( a_k = k^k \)), the series diverges.
- If \( q = 1 \), the series may either converge or diverge.

Proof. The case when the limit \( q \) does not exist is obvious by the contraposition to proposition 114.
Suppose that the limit exists. Therefore, there exists an index \( K \) such that
\[ \sqrt[k]{a_k} \leq q \quad \forall k \geq K. \]
Thus, we have the inequality
\[ a_k \leq q^k \quad \forall k \geq K. \]
The statement of the theorem now follows from comparison (proposition 64) with the geometric series (534).

Proposition 66 (d’Alambert’s ratio test). Consider the nonnegative series eq. (8). Put
\[ q := \limsup_{k \to \infty} \frac{a_{k+1}}{a_k}, \]
where \( q = \infty \) if the limit does not exist. Then
- If \( q < 1 \), the series converges.
- If there exists an index \( k_0 \) such that \( \frac{a_{k+1}}{a_k} \geq 1 \) for all \( k \geq k_0 \), the series diverges.
- If the limit does not exist (e.g. if \( a_k = k! \)), the series diverges.

Proof. All cases except for \( q < 1 \) are obvious by the contraposition to proposition 114.
Suppose that the limit exists. Therefore, there exists an index \( k_0 \) such that
\[ a_{k+1} \leq qa_k \quad \forall k \geq k_0. \]
Thus,
\[ a_{k_0+m} \leq q^m a_{k_0} \quad \forall m \geq \mathbb{Z}^{\geq 0}. \]
Convergence now follows from comparison (proposition 64) of the geometric series (534) with the subseries of eq. (8) obtained by trimming the first \( k_0 \) elements.
Proposition 67. The values of $q$ in proposition 65 (Cauchy’s root test) and in proposition 66 (d’Alambert’s ratio test) are equal.

Proof. If we assume that they are not equal, then the same series would have to be convergent and divergent simultaneously in some region. □

Definition 68. Series of the form
\[ \pm \sum_{k=0}^{\infty} (-1)^k a_k, \]  
(10)
where all $a_k$, $k = 0, 1, \ldots$ are nonnegative, are called alternating.

Proposition 69 (Leibniz’ alternating series test). Consider the alternating series definition 68. If the sequence of terms $\{a_k\}_{k=0}^{\infty}$ decreases monotonically and if $\lim_{k \to \infty} a_k = 0$, then the series converges.

Theorem 70. Corollary 118 (Weierstrass’ series criterion) is a necessary condition for nonnegative real functions.
2.6. Riemann integration

Definition 71. The concept of a partition of a nonempty real closed interval \([a, b]\) is the base for defining Riemann-style integrals.

(a) A **Riemann partition** of \([a, b]\) is a set

\[ \Delta := \{x_0, \ldots, x_n\} \subseteq [a, b] \]

that satisfies

\[ a = x_0 < x_1 < \ldots < x_n = b. \]

For brevity, we write

\[ \Delta : a = x_0 < x_1 < \ldots < x_n = b. \]  \hfill (11)

We denote the set of all partitions of \([a, b]\) by \(\text{part}([a, b])\).

(b) The partition

\[ \Gamma : a = y_0 < y_1 < \ldots < y_m = b \]

is called a **refinement** of the partition (11) if we have the set inclusion

\[ \{x_0, x_1, \ldots, x_n\} \subseteq \{y_0, y_1, \ldots, y_m\}. \]  \hfill (12)

In this case, we “split” \(\Gamma\) into chains such that, for each \(k = 1, 2, \ldots, n\),

\[ y_{k,j} := \begin{cases} x_{k-1}, & j = 0, \\ x_k, & j = p_k, \\ \text{\(j\)-th point of} \{y_0, \ldots, y_m\} \cap [x_{k-1}, x_k], & 0 < j < p_k. \end{cases} \]  \hfill (13)

(c) Finally, the **diameter** of the partition (11) is defined as

\[ \text{diam}(\Delta) := \max_{1 \leq k \leq n} (x_k - x_{k-1}). \]  \hfill (14)

(d) We can make the set \(\text{part}([a, b])\) of all Riemann partitions of \([a, b]\) into a directed set using two common approaches:

(i) Put \(\Delta \leq_R \Gamma\) if and only if \(\Gamma\) is a refinement of \(\Delta\). This actually makes \((\text{part}([a, b]), \leq_R)\) a partially ordered set.

(ii) Put \(\Delta \leq_D \Gamma\) if and only if \(\text{diam}(\Gamma) \leq \text{diam}(\Delta)\).

(e) A **tagged partition** of \([a, b]\) is a partition (11) of \([a, b]\) along with a choice of a tag \(\xi_k\) for each closed interval \([x_{k-1}, x_k]\), \(k = 1, \ldots, n\). By putting \(\Xi := \{\xi_k\}_{k=1}^n\), we can define a tagged partition as the tuple \((\Delta, \Xi)\). For brevity, we write

\[ \Delta : a = x_0 < x_1 < \ldots < x_n = b \]

\[ \Xi : \xi_k \in [x_{k-1}, x_k], k = 1, \ldots, n. \]  \hfill (15)
We denote the set of all tagged partitions of \([a,b]\) by \(\text{tpart}([a,b])\). We introduce an order on \(\text{tpart}([a,b])\) by putting
\[
(\Delta, \Xi) \preceq_R (\Gamma, H) \quad \text{if and only if} \quad \Delta \preceq_R H
\]
and analogously for \(\preceq_D\). Note that \(\preceq_R\) is not a partial order in \(\text{tpart}([a,b])\) unlike in \(\text{part}([a,b])\).

**Remark 72.** Note that (11) is not a partition in the sense of definition 959, however the set of intervals
\[
\left\{[x_0, x_1), [x_1, x_2), \ldots, [x_{n-2}, x_{n-1}), [x_{n-1}, x_n]\right\}
\]
is a set-theoretic partition of \([a,b]\). Conversely, every finite set-theoretic partition of \([a,b]\) gives rise to a Riemann partition in the sense of definition 71 (a).

**Definition 73.** Let \(X\) be a real Hausdorff topological vector space. Fix a function \(f : [a,b] \to X\).

The **Riemann sum** of \(f\) corresponding to the tagged partition (15) is defined as
\[
S(f, \Delta, \Xi) := \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}).
\]

Consider the net
\[
\{S(f, \Delta, \Xi)\}_{(\Delta, \Xi) \in \text{tpart}([a,b])}
\]  
(16)

Both orders definition 71 (d i) and definition 71 (d ii) on \(\text{tpart}([a,b])\) provide equivalent convergence for Riemann sums. If the limit exists, \(f\) is said to be **Riemann integrable** in \([a,b]\). We call the limit the **Riemann integral** of \(f\) and denote it by
\[
\int_{a}^{b} f(x) \, dx.
\]  
(17)

**Proof.**

**Proof that 71 (d i) implies 71 (d ii).** Let \(I\) be the limit (17) with respect to \(\preceq_R\). Fix a neighborhood \(U\) of 0. Since (16) is eventually in \(I + U\), there exists a tagged partition
\[
\Delta_0 : a = x_0^{(0)} < x_1^{(0)} < \ldots < x_n^{(0)} = b
\]
\[
\Xi_0 : \xi_k^{(0)} \in [x_{k-1}^{(0)}, x_k^{(0)}], k = 1, \ldots, n_0.
\]  
(18)
such that \(S(f, \Gamma, H) \in I + U\) if \(\Gamma\) is a refinement of \(\Delta_0\).

Note that \(f\) is **bounded**. Indeed, if it is unbounded on \([a,b]\), then there exists a refinement \((\Gamma, H)\) of \((\Delta_0, \Xi_0)\) such that
\[
S(f, \Gamma, H) - I \notin U.
\]

But this contradicts our choice of \(\Delta_0\). Therefore, \(f\) is bounded and there exists a bounded neighborhood \(V_0\) of 0 such that \(f([a,b]) \subseteq V_0\) and hence \(f(x) - f(y) \in V := V_0 - V_0\) for all \(x, y \in [a,b]\).
Let \( v > 0 \) be such that \( V \subseteq vU \).

Let \((\Delta, \Xi)\) be a tagged partition such that \( \text{diam}(\Delta) \leq \text{diam}(\Delta_0) \).

We introduce another partition \( \Gamma := \Delta \cup \Delta_0 \). Since \( \Gamma \) is a refinement of \( \Delta_0 \), we can use a splitting similar to (13) such that

\[
S(f, \Delta, \Xi) = \sum_{k=1}^{n_0} \sum_{j=1}^{p_k} f(\xi_j) \big( y_{k,j} - y_{k,j-1} \big).
\]

(19)

Denote by \( \xi_{k,j} \) the largest tag in \( \Xi \) such that \( \xi_{k,j} \leq y_{k,j} \). Thus,

\[
S(f, \Delta, \Xi) = \sum_{k=1}^{n_0} \sum_{j=1}^{p_k} f(\xi_{k,j}) \big( y_{k,j} - y_{k,j-1} \big).
\]

For every \( k = 1, \ldots, n \) and every \( j = 0, \ldots, p_k \), choose an arbitrary tag

\( H : \eta_{k,j} \in [y_{k,j-1}, y_{k,j}] \).

Then we have

\[
S(f, \Delta, \Xi) - I = S(f, \Delta, \Xi) - S(f, \Gamma, H) + S(f, \Gamma, H) - I \in \sum_{\in U} \sum_{k=1}^{n_0} \sum_{j=1}^{p_k} \big( f(\xi_{k,j}) - f(\eta_{k,j}) \big) \big( y_{k,j} - y_{k,j-1} \big) + U \subseteq \]

\[
\subseteq V \cdot \sum_{k=1}^{n_0} \sum_{j=1}^{p_k} \big( y_{k,j} - y_{k,j-1} \big) + U \subseteq \]

\[
\subseteq \text{diam}(\Delta) \cdot n_0 \cdot V + U \subseteq \]

\[
\subseteq (\text{diam}(\Delta) \cdot n_0 \cdot v + 1)U.
\]

Let \((\Delta_1, \Xi_1)\) be a tagged partition of \([a, b]\) such that \( \text{diam}(\Delta_1) \leq \min \left\{ \text{diam}(\Delta_0), \frac{1}{vn_0} \right\} \). It follows that

\[
S(f, \Delta_1, \Xi_1) - I \subseteq 2U.
\]

(20)

Until now, \( U \) was fixed. Given any neighborhood \( W \) of 0, we need to choose a neighborhood \( U \) of 0 and a corresponding partition \( \Delta_1 \) such that (20) holds. Then, whenever \( \text{diam}(\Delta) \leq \text{diam}(\Delta_1) \), we have

\[
S(f, \Delta, \Xi) - I \subseteq 2U \subseteq W.
\]

This finishes the proof.

**Proof that 71 (d ii) implies 71 (d i).** Note that if \( \Gamma \) is a refinement of \( \Delta \), clearly \( \text{diam}(\Gamma) \leq \text{diam}(\Delta) \). Therefore, if the net (16) with respect to \( \leq \mathcal{D} \) is eventually in some open set \( U \), the corresponding net with respect to \( \leq \mathcal{R} \) is also eventually in \( U \). This finishes the proof. \( \Box \)
Corollary 74. A Riemann-integrable function is bounded.

Proof. Proven in definition 73.

Definition 75. Let \((X, \rho)\) be a Frechet space. Fix a function \(f : [a, b] \to X\). Similarly to definition 73, choose any of the orderings definition 71 (d i) and definition 71 (d ii) on the set of all untagged Riemann partitions \(\text{part}([a, b])\).

For each partition (11), we define its oscillation via the function oscillation of \(f\)

\[
\omega(f, \Delta) := \sum_{k=1}^{n} \omega(f, [x_{k-1}, x_k])(x_k - x_{k-1}).
\]

Consider the net

\[
\{\omega(f, \Delta)\}_{\Delta \in \text{part}([a,b])}
\]

If this net converges to zero, we say that \(f\) is Darboux integrable.

Proposition 76. In a Banach space, Darboux integrability implies Riemann integrability.

Proof. We will show that the net (16) is fundamental. Fix \(\varepsilon > 0\). Since \(f\) is Darboux integrable, there exists an untagged partition \(\Delta_0\) such that, if \(\Delta\) is a refinement of \(\Delta_0\), we have

\[
\omega(f, \Delta) < \varepsilon.
\]

Let \(\Delta\) be a refinement of \(\Delta_0\) and \(\Gamma\) be a refinement of \(\Delta\). Assume that the points of \(\Gamma\) are split as in (13). Choose arbitrary tags \(\Xi = \{\xi_k\}_{k=1}^{n}\) for \(\Delta\) and \(H = \{\eta_{k,j}\}_{k=1, j=1}^{n,p_k}\) for \(\Gamma\). For the corresponding Riemann sums, we have

\[
\|S(f, \Delta, \Xi) - S(f, \Gamma, H)\| = \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}) - \sum_{k=1}^{n} \sum_{j=1}^{p_k} f(\eta_{k,j})(y_{k,j} - y_{k,j-1}) \leq \omega(f, \Delta) < \varepsilon.
\]

Therefore, the net (16) is fundamental and, since \(X\) is complete, the net converges to a limit. \(\square\)
Definition 77. Fix a real-valued function $f : [a, b] \to \mathbb{R}$. The upper Darboux sum corresponding to the partition (11) is defined as

$$
\bar{S}(f, \Delta) := \sum_{k=1}^{n} (x_k - x_{k-1}) \sup_{\xi \in [x_{k-1}, x_k]} f(\xi).
$$

The lower Darboux sum is defined as

$$
\underline{S}(f, \Delta) := \sum_{k=1}^{n} (x_k - x_{k-1}) \inf_{\xi \in [x_{k-1}, x_k]} f(\xi).
$$

If the nets

$$
\{\bar{S}(f, \Delta)\}_{\Delta \in \text{part}([a, b])} \quad \{\underline{S}(f, \Delta)\}_{\Delta \in \text{part}([a, b])}
$$

have a common limit, we call this limit the Darboux integral of $f$ and, analogously to definition 73, we denote it by

$$
\int_{a}^{b} f(x)dx.
$$

This notation is justified by proposition 79.

Proposition 78. A real-valued function $f : [a, b] \to \mathbb{R}$ is Darboux integrable if and only if it has a Darboux integral.

Proof. Note that, given the partition (11), we have

$$
\bar{S}(f, \Delta) - \underline{S}(f, \Delta) = \sum_{k=1}^{n} (x_k - x_{k-1}) \left[ \sup_{\xi \in [x_{k-1}, x_k]} f(\xi) - \inf_{\eta \in [x_{k-1}, x_k]} f(\eta) \right] = \\
= \sum_{k=1}^{n} (x_k - x_{k-1}) \left[ \sup_{\xi \in [x_{k-1}, x_k]} f(\xi) + \sup_{\eta \in [x_{k-1}, x_k]} -f(\eta) \right] = \\
= \sum_{k=1}^{n} (x_k - x_{k-1}) \sup \{f(\xi) - f(\eta) : \xi, \eta \in [x_{k-1}, x_k]\} = \\
= \sum_{k=1}^{n} (x_k - x_{k-1}) \sup \{|f(\xi) - f(\eta)| : \xi, \eta \in [x_{k-1}, x_k]\} = \\
= \omega(f, \Delta).
$$

Therefore, the nets (23) converge to a common limit if and only if $\omega(f, \Delta) \to 0$. This finishes the proof.

Proposition 79. A real-valued function $f : [a, b] \to \mathbb{R}$ has a Darboux integral if and only if it has a Riemann integral. Furthermore, the two integrals are equal.

Proof. Fix $\varepsilon > 0$. 

46
**Proof that 77 implies 73.** Denote by $I_D$ the Darboux integral of $f$. Then there exists a partition $\Delta_0$ of $[a, b]$ such that for any refinement (11) of $\Delta_0$ we have

$$S(f, \Delta) - S(f, \Delta) < \frac{\varepsilon}{2}.$$

In particular, $I_D - S(f, \Delta) < \frac{\varepsilon}{2}$.

Let $\Xi := \{\xi_k\}_{k=1}^n$ be tags for $\Delta$. Then

$$|S(f, \Delta, \Xi) - I_D| \leq |S(f, \Delta, \Xi) - S(f, \Delta)| - |S(f, \Delta) - I| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Therefore, $I_D$ is also a Riemann integral for $f$.

**Proof that 73 implies 77.** Denote by $I_R$ the Riemann integral of $f$. Then there exists a partition (18) such that for any partition (15) with $\text{diam}(\Delta) \leq \text{diam}(\Delta_0)$, we have

$$|S(f, \Delta, \Xi) - I_R| < \frac{\varepsilon}{2}.$$

Since (74) is bounded, there exists a constant $M > 0$ such that $|f(\xi) - f(\eta)| < M$ for any $\xi, \eta \in [a, b]$.

Using an analogous to (19) splitting for the refinement $\Gamma := \Delta \cup \Delta_0$ of $\Delta_0$, we obtain

$$\overline{S}(f, \Gamma) - S(f, \Gamma, H) = \sum_{k=1}^{n_0} \sum_{k=1}^{p_k} \left[ \sup_{\xi \in [y_{k,j-1}, y_{k,j}]} f(\eta) - f(\eta_{k,j}) \right] (y_{k,j} - y_{k,j-1}) \leq M \sum_{k=1}^{n_0} \sum_{k=1}^{p_k} (y_{k,j} - y_{k,j-1}) \leq M \cdot n_0 \cdot \text{diam(}\Gamma).$$

By choosing a tagged partition $(\Delta_1, \Xi_1)$ with $\text{diam}(\Delta_1) < \min\{\text{diam(}\Delta_0), \frac{\varepsilon}{2Mn_0}\}$, we obtain

$$\overline{S}(f, \Delta_1) - S(f, \Delta_1, \Xi) < \frac{\varepsilon}{2}.$$

Therefore, whenever $\text{diam}(\Delta) \leq \text{diam}(\Delta_1)$,

$$\overline{S}(f, \Delta) - I_R = \overline{S}(f, \Delta) - S(f, \Delta, \Xi) + S(f, \Delta, \Xi) - I_R < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus, the net $\{\overline{S}(f, \Delta)\}_{\Delta \in \text{part([a,b])}}$ of upper Darboux sums converges to $I$. We can analogously show that the lower Darboux sums also converge to $I_R$. Hence, $I_R$ is the Darboux integral of $f$. \qed
Proposition 80. In a Frechet space $(X, \rho)$, continuous functions $f : [a, b] \to X$ are Darboux integrable.

Proof. Fix $\delta > 0$. Let (11) be a partition of $[a, b]$ such that $\text{diam} (\Delta) < \delta$. We have

$$\omega(f, \Delta) = \sum_{k=1}^{n} \omega(f, [x_{k-1}, x_k])(x_k - x_{k-1}) \leq \sum_{k=1}^{n} \omega(f, \text{diam} (\Delta)) \text{diam} (\Delta) < n \omega(f, \delta) \delta.$$ 

Now fix $\varepsilon > 0$. A continuous function on a compact interval is uniformly continuous. By proposition 392 (a), there exists $\delta_0 > 0$ such that $\omega(f, \delta_0) < \varepsilon$. It is then enough to choose

$$\delta := \frac{\delta_0}{n \varepsilon}$$

to obtain

$$\omega(f, \Delta) < n \delta \omega(f, \delta) = \delta_0 \frac{\omega(f, \delta)}{\varepsilon} \leq \delta_0 \frac{\omega(f, \delta_0)}{\varepsilon} < \varepsilon.$$ 

Therefore, the same inequality holds for all partitions with diameters smaller than $\delta$, which implies that $f$ is Darboux integrable. \hfill \Box

Proposition 81. Let $f : [a, b] \to \mathbb{R}^n$ be a function and let $f_k, k = 1, \ldots, n$ be its components. We have that $f$ is integrable if and only if $f_k$ is integrable for $k = 1, \ldots, n$. Furthermore,

$$\int_a^b \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} dx = \begin{pmatrix} \int_a^b f_1(x) dx \\ \vdots \\ \int_a^b f_n(x) dx \end{pmatrix}.$$  

(24)

Proof.

Proof of sufficiency. Let $f$ be integrable and let $I = (I_1, \ldots, I_n)^T$ be the value of the integral. Fix $\varepsilon > 0$ and let $(\Delta, \Xi)$ be a tagged partition such that

$$||I - S(f, \Delta, \Xi)|| < \varepsilon.$$ 

Then for any $k = 1, \ldots, n$ we have

$$||I - S(f, \Delta, \Xi)||^2 = \sum_{m=1}^{n} |I_m - S(f_m, \Delta, \Xi)|^2 \geq |I_k - S(f_k, \Delta, \Xi)|^2,$$

hence

$$|I_k - S(f_k, \Delta, \Xi)| < \varepsilon.$$ 

Therefore, $f_k$ is integrable and

$$\int_a^b f_k(x) dx = I_k.$$
**Proof of necessity.** Let $f_k$ be integrable for $k = 1, \ldots, n$ with value $I_k$. Put $I := (I_1, \ldots, I_n)^T$. Fix $\delta > 0$ and let $(\Delta_k, \Xi_k)$ be a partition such that

$$|I_k - S(f_k, \Delta_k, \Xi_k)| < \delta$$

Let $\Gamma := \bigcup_{k=1}^n \Delta_k$ and let $H$ be tags for $\Gamma$. Since $\text{diam}(\Gamma) \leq \text{diam}(\Delta_k)$ and since $f_k$ is integrable, we have

$$|I_k - S(f_k, \Gamma, H)| < \delta \quad \forall k = 1, \ldots, n.$$ 

We have

$$\|I - S(f, \Gamma, H)\| = \sqrt{\sum_{m=1}^n |I_m - S(f_m, \Gamma, H)|^2} < \delta \sqrt{n}.$$ 

Therefore, given $\epsilon > 0$, it is enough to choose $\delta := \frac{\epsilon}{\sqrt{n}}$ to obtain a tagged partition $(\Gamma_0, H_0)$, so that for $(\Gamma, H)$ with $\text{diam}(\Gamma) < \text{diam}(\Gamma_0)$ we have

$$\|I - S(f, \Gamma, H)\| < \epsilon.$$ 

This proves integrability of $f$ and (24).
2.7. Line integrals

**Definition 82.** Let $X$ be a Banach space and let $\gamma : [a, b] \to X$ be a $1$-in order to find the curve’s length, we use the following procedure:

- Fix a nonnegative integer $n$.
- Choose $n - 1$ points $c_1, \ldots, c_{n-1}$ in $[0, 1]$ and order them ascendingly. Define $c_0 := 0$ and $c_n := 1$. We will call the tuple $c := (c_0, \ldots, c_n)$ a **partition** of $[0, 1]$ because it partitions $[0, 1]$ into the subintervals $[c_{k-1}, c_k], k = 1, \ldots, n$.

Note that this choice does not actually require the axiom of choice since we will universally quantify all partitions.
- Use the partition $c$ to build linear approximation to $\gamma$ using the line segments $[\gamma(c_{k-1}), \gamma(c_k)], k = 1, \ldots, n$.
- Find the total length of the approximation as

$$\text{len}_c(\gamma) := \sum_{k=1}^{n} \|\gamma(c_k) - \gamma(c_{k-1})\|.$$ 

- Build a **directed set** of all partitions by introducing an order that depends only the size of the tuples (i.e. we declare all partitions with the same size as equal).
- Using the defined directed set, build a **net** that assigns to each partition the length $\text{len}_c(\gamma)$. The limit of this net, if it exists, is called the **length** of the curve $\gamma$ and is denoted by $\text{len}(\gamma)$.

If the curve $\gamma$ has a length, it is called **rectifiable**.

**Proposition 83.** For a differentiable parametric curve $\gamma : [a, b] \to X$ we have

$$\text{len}(\gamma) := \int_a^b \|\gamma'(x)\| \, dx.$$ 

**Proof.** By the mean value theorem, when constructing the length in definition 82, for each $k = 1, \ldots, n$ there exists a point $\xi_k \in [c_{k-1}, c_k]$ such that

$$\gamma(c_k) - \gamma(c_{k-1}) = \gamma'(\xi_k)(c_k - c_{k-1}).$$

Therefore,

$$\text{len}_n(\gamma) = \sum_{k=1}^{n+1} \|\gamma(c_k) - \gamma(c_{k-1})\| = \sum_{k=1}^{n+1} \|\gamma'(\xi_k)\|(c_k - c_{k-1}) \to \int_a^b \|\gamma'(x)\| \, dx.$$ 

\[\Box\]
Corollary 84. The length of the graph of a differentiable function $f : [a, b] \to \mathbb{R}$, if it exists, is given by

$$\text{len}(\text{gph}(f)) := \int_a^b \sqrt{1 + f'(x)}\,dx.$$  

Proof. Apply proposition 83 for the parametric curve $\gamma(x) := \text{gph}(y^+(x)) = (x, f(x)).$
2.8. Total variation

**Definition 85.** The most common generalization of the Riemann integral is the Riemann-Stieltjes integral. It is not as well-behaved, hence we will give the most general definition and not attempt to prove equivalences.

Let $X$ be a real Hausdorff topological vector space. Fix two functions $f, \alpha : [a, b] \to X$.

The **Riemann sum** of $f$ with respect to $\alpha$ corresponding to the tagged partition (15) is defined as

$$S(f, \alpha, \Delta, \Xi) := \sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})).$$

The limit of the net

$$\{S(f, \alpha, \Delta, \Xi)_{(\Delta, \Xi) \in \text{tpart}([a,b])}\}, \tag{25}$$

if it exists, is called the **Riemann-Stieltjes integral** of $f$ with respect to $\alpha$ and is denoted by

$$\int_{a}^{b} f(x) d\alpha(x).$$
2.9. Nonsmooth derivatives

Remark 86. Nonsmooth analysis studies generalized differentiability for functions which are not necessarily differentiable. The generalized derivatives (see section 2.9 (Nonsmooth derivatives)) are not linear, which motivates the study of subdifferentials (see section 2.11 (Subdifferentials)).

Both optimization in Euclidean spaces and infinite-dimensional optimization studies (not necessarily linear) real-valued functionals. Hence, we are only concerned with studying real-valued topological vector spaces.

Remark 87. Unlike in definition 204, we do not introduce terminology for differentiability because actual differentiability refers to linear approximations of \( f : U \rightarrow Y \) with some consistency properties. We will say that “\( f \) has a Clarke derivative at \( x_0 \) in the direction \( h \)” rather than “\( f \) is Clarke differentiable at \( x_0 \) in the direction \( h \)”.

Definition 88. Let \( X \) be a real Hausdorff topological vector spaces and let \( U \subset X \) be an open set.

We fix a point \( x_0 \in U \) and a direction \( h \in X \).

(a) Definition 204 already introduced directional derivatives. Here we introduce a special notation for them:

\[
D^+_h f(x_0) = f'_+(x_0)(h) := \lim_{t \downarrow 0} \frac{f(x_0 + th) - f(x_0)}{t}.
\]

(b) [Cla13, definition 11.18] The upper (resp. lower) Dini derivative is defined as

\[
\overline{D}_h f(x_0) = \overline{f}'(x_0)(h) := \limsup_{t \downarrow 0} \frac{f(x_0 + th) - f(x_0)}{t}
\]

\[
\underline{D}_h f(x_0) = \underline{f}'(x_0)(h) := \liminf_{t \downarrow 0} \frac{f(x_0 + th) - f(x_0)}{t}
\]

Dini derivatives are useful when the difference quotients are bounded, but do not have a limit.

(c) [Cla13, section 10.1] The generalized Clarke derivative is defined as

\[
D^\circ_h f(x_0) = f^\circ(x_0)(h) := \limsup_{y \rightarrow x_0 \atop t \downarrow 0} \frac{f(y + th) - f(y)}{t}.
\]

Refer to section 2.12 (Clarke generalized gradients) for their usefulness.
2.10. Convex functions

Let $X$ be a Hausdorff topological vector space and $D$ be a convex subset of $X$.

**Definition 89.** A function $f : D \to \mathbb{R}$ is called **convex** if any of the following equivalent conditions hold:

(a) For any two points $x, y \in D$ and any $t \in [0,1]$ we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

(b) The epigraph

$$\text{epi } f := \{(x, a) \in X \times \mathbb{R} : f(x) \leq a\}$$

is convex.

If $-g$ is convex for some function $g : D \to \mathbb{R}$, we call $g$ **concave**.

Note that definitions do not require any topological structure on $X$. Most of their properties, however, require a topology.

**Proof.** Let $x, y \in D$ and let $t \in [0,1]$.

**Proof that 89 (a) implies 89 (b).** Let $\text{epi } f$ be a convex set. Obviously $(x, f(x)) \in D$ and $(y, f(y)) \in D$. By the convexity of $\text{epi } f$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Thus, $f$ is a convex function.

**Proof that 89 (b) implies 89 (a).** Let $f$ be convex. Let $a \geq f(x)$ and $b \geq f(y)$, so that $(x, a) \in \text{epi } f$ and $(y, b) \in \text{epi } f$. Hence,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \leq ta + (1-t)b,$$

which implies that

$$(tx + (1-t)y, ta + (1-t)b) \in \text{epi } f.$$ 

Thus, $\text{epi } f$ is a convex set.

**Definition 90.** **Affine functions** $f : X \to \mathbb{R}$ are simultaneously convex and concave.

**Proposition 91.** For any convex function $f$ and any $x \in D$, the set $\partial f(x)$ is convex and weak* closed.

**Proof.** Fix $x \in D$. If $\partial f(x)$ is empty, then the theorem is trivially true.

Suppose it is nonempty and $y^*, z^* \in \partial f(x)$. For any $x \in D$ we then have

$$\langle y^*, x - x \rangle \leq f(x) - f(x),$$

$$\langle z^*, x - x \rangle \leq f(x) - f(x).$$

54
Fix $t \in [0, 1]$ and $x \in D$. It follows that

$$
\langle ty^* + (1-t)z^*, x-x \rangle = t\langle y^*, x-x \rangle + (1-t)\langle z^*, x-x \rangle \\
\leq t[f(x) - f(x)] + (1-t)[f(x) - f(x)] = f(x) - f(x),
$$

thus $ty^* + (1-t)z^* \in \partial f(x)$ and hence $\partial f(x)$ is convex.

To prove weak*-closedness, we consider the decomposition

$$
\partial f(x) = \{x^* \in E^* : \forall x \in D, \langle x^*, x-x \rangle \leq f(x) - f(x)\} =
\bigcap_{x \in D} L(x)^{-1}(-\infty, f(x) - f(x)),
$$

where

$$
L : E \to E^{**} \\
L(x)(x^*) = \langle x^*, x-x \rangle.
$$

For each $x \in E$, the functionals $L(x)$ are weak*-to-weak continuous because the image $L(E) \subseteq E^{**}$ is isometrically isomorphic to a translation of $E$. Hence, the preimage $L(x)^{-1}(-\infty, f(x) - f(x)]$ is closed and $\partial f(x)$ is weak*-closed as the intersection of weak*-closed sets.

**Lemma 92.** For every point $x \in X$ and every direction $h \in S_X$ the difference quotient is a monotone function of $t > 0$, i.e. for $0 < s < t$

$$
\frac{f(x + sh) - f(x)}{s} \leq \frac{f(x + th) - f(x)}{t}
$$

**Proof.**

$$
\frac{f(x + sh) - f(x)}{s} = \frac{t}{s} \frac{f(x + \frac{s}{t}h) - f(x)}{\frac{s}{t}} \leq \frac{t}{s} \frac{f \left( \frac{s}{t} (x + th) + \left( 1 - \frac{s}{t} \right)x \right) - f(x)}{\frac{s}{t} - \frac{s}{t} f(x) - f(x)} \leq \frac{t}{s} \frac{f(x + th) - f(x)}{\frac{s}{t}} \leq \frac{f(x + th) - f(x)}{t}
$$

**Proposition 93.** For every point $x \in X$ and every direction $h \in S_X$ the one-sided derivative $f'_+(x)(h)$ exists.

**Proof.** We use the convexity of $f$ to obtain

$$
f(x) = f \left( x + \frac{th}{2} - \frac{th}{2} \right) \leq \frac{f(x + th) + f(x - th)}{2},
$$
\[
0 \leq [f(x - th) - f(x)] + [f(x + th) - f(x)], \\
-\frac{f(x + t(-h)) - f(x)}{t} \leq \frac{f(x + th) - f(x)}{t},
\]

thus the difference quotient in \( f'_+(x)(h) \) is bounded below by the difference quotient for 
\( -f'_+(x)(-h) \).

Lemma 92 implies that the right difference quotient is non-increasing, thus both limits exist and
\[
-f'_+(x)(-h) \leq f'_+(x)(h).
\]

\[\square\]

**Proposition 94.** For every point \( x \in X \) and every direction \( h \in S_X \) the one-sided derivative \( f'_+(x)(h) \) is a sublinear functional.

**Proof.**

**Proof of 183 (a).** It follows directly from
\[
\frac{f(x + t(a + b)) - f(x)}{t} = \frac{f(\frac{1}{2}(x + 2ta) + \frac{1}{2}(x + 2tb)) - f(x)}{t} \leq \frac{\frac{1}{2}f(x + 2ta) + \frac{1}{2}f(x + 2tb) - f(x)}{t} = \frac{f(x + 2ta) - f(x)}{2t} + \frac{f(x + 2tb) - f(x)}{2t}.
\]

**Proof of 183 (b).** For \( \lambda > 0 \) the equality \( f'_+(x)(\lambda h) = \lambda f'_+(x)(h) \) follows from
\[
\frac{f(x + t\lambda h) - f(x)}{t} = \lambda \frac{f(x + t\lambda h) - f(x)}{t\lambda}
\]

\[\square\]

**Corollary 95.**
\[-f_{+}'(x)(-h) \leq f'_+(x)(h)\]

**Proof.**
\[0 = f'_+(x)(h + (-h)) \leq f'_+(x)(h) + f'_+(x)(-h)\]

\[\square\]

**Proposition 96.** The continuous function \( f : D \to X \) is convex if and only if its subdifferential \( \partial f(x) \) (see definition 103 (a)) is nonempty for every \( x \) in \( D \).

**Proof.** TODO: Prove.

**Proposition 97.** For every direction \( h \in S_X \), we have that
\[f'_+(x)(h) = \max\{\langle x^*, h \rangle : x^* \in \partial f(x)\}\].
**Theorem 98.** If $f$ is continuous and if the subdifferential $\partial f(x)$ at $x \in X$ is a singleton with element $x^*$, then $f$ is Gateaux differentiable at $x$ and $f'_G(x) = x^*$.

**Proof.** Let $h \in S_X$ be arbitrary. **Proposition 93** implies that the one-sided derivatives $f'_+(x)(-h)$ and $f'_+(x)(h)$ exist and

$$-f'_+(x)(-h) \leq f'_+(x)(h).$$

Assume that $f$ is not Gateaux differentiable at $x$, i.e. for some $h_0 \in X$, we have a strict inequality. Then by **proposition 97**

$$\min\{(x^*, h_0) : x^* \in \partial f(x)\} = -\max\{(x^*, -h_0) : x^* \in \partial f(x)\} < -f'_+(x)(h_0) = \max\{(x^*, h_0) : x^* \in \partial f(x)\},$$

which implies that there is more that one functional $x^* \in \partial_C f(x)$. This contradicts the assumption of the theorem.

Thus, $f$ is Gateaux differentiable at $x$. \qed

**Theorem 99.** In $\mathbb{R}^n$, the existence of the partial derivatives at $x$ for a continuous convex function $f : D \rightarrow \mathbb{R}$ at a point $x \in D$ implies Gateaux differentiability.

**Proof.** Let $D \subseteq \mathbb{R}^n$ be an open and convex set and let $f : D \rightarrow \mathbb{R}$ be continuous and convex. Then $f'_+(x)$ exists everywhere by **proposition 93** and is a subdifferential functional by **proposition 94**.

Let $e_1, \ldots, e_n$ be the canonical basis for $\mathbb{R}^n$.

The partial derivatives

$$\frac{\partial f}{\partial x_i}(x) := \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t} = f'_+(x)(e_i)$$

exist, hence the projections of $f'_+(x)$ along the coordinate axes are linear.

Define line linear functional

$$l(h) := \sum_{i=1}^{n} h_i \langle \frac{\partial f}{\partial x_i}(x), h \rangle,$$

where $h_1, \ldots, h_n$ are the coordinates of $h$ along $e_1, \ldots, e_n$.

We will show that $l \equiv f'_+(x)$. Fix $h \in S_X$. We have

$$f'_+(x)(h) = f'_+(x) \left( \sum_{i=1}^{n} h_i e_i \right) \overset{\text{sublinearity}}{\leq} \sum_{i=1}^{n} f'_+(x)(h_i e_i) \overset{\text{linearity along } e_i}{=} \sum_{i=1}^{n} h_i f'_+(x)(e_i) = \sum_{i=1}^{n} h_i \langle \frac{\partial f}{\partial x_i}(x), h \rangle. \quad (26)$$

57
Thus,

$$\langle l, h \rangle = -\langle l, -h \rangle \leq -f'_+(x)(-h) \leq f'_+(x)(h) \leq \langle l, h \rangle,$$

i.e. $f'_+(x)(h) = \langle l, h \rangle$ for all $h \in S_X$, hence $f'_+(x)$ is a linear functional and $f$ is Gateaux differentiable at $x$. 

\[ \square \]

**Theorem 100.** In $\mathbb{R}^n$, Gateaux differentiability of a continuous convex function $f : D \to \mathbb{R}$ at a point $x \in D$ implies Frechet differentiability.

**Proof.** Since $f$ is Gateaux differentiable (definition 204 (b)) at $x$, the derivative $f'(x) = f'_+(x)$ is linear.

Because $f$ is continuous and convex, it is locally Lipschitz with constant $L$ in some $\delta$-ball with center $x$.

Suppose that $f$ is not Frechet differentiable at $x$. Inverting the condition in definition 204 (c), we obtain that there exist $\varepsilon > 0$ and a sequence $\{h_n\} \subseteq B(x, \delta) \setminus \{0\}$ such that $\|h_n\| \to 0$ and yet for all $n \in \mathbb{Z}_{>0}$,

$$|f(x + h_n) - f(x) - \langle f'(x), h_n \rangle| > \varepsilon\|h_n\|. \quad (27)$$

Define

$$t_n := \|h_n\| \quad u_n := \frac{h_n}{\|h_n\|}. \quad (28)$$

Fix $\delta > 0$. Because of the Gateaux differentiability of $f$ at $x$, we can pick $k_0$ such that

$$\left| \frac{f(x + t_{n_k}u_{n_k}) - f(x)}{t_{n_k}} - \langle f'(x), u_{n_k} \rangle \right| < \delta.$$ 

Because $\{u_{n_k}\}_{k}$ converges to $u_0$, we can choose $k_1$ such that

$$\|u_0 - u_{n_k}\| < \delta.$$
Thus, for $k > \max\{k_0, k_1\}$, eq. (28) is bounded by
\[
\left| \frac{f(x + t_n u_{n_k}) - f(x)}{t_{n_k}} - \langle f'(x), u_{n_k} \rangle \right| \leq (L + 1 + \|f'(x)\|)\delta.
\]
It suffices to choose $\delta > 0$, so that
\[
\delta < \frac{1}{L + 1 + \|f'(x)\|}
\]
in order to have, for $k > \max\{k_0, k_1\}$,
\[
\left| \frac{f(x + t_n u_{n_k}) - f(x)}{t_{n_k}} - \langle f'(x), u_{n_k} \rangle \right| < \varepsilon.
\]
But this contradicts eq. (27), hence $f$ is Frechet differentiable at $x$.

**Corollary 101.** In $\mathbb{R}^n$, the existence of the partial derivatives at $x$ for a continuous convex function $f : D \to \mathbb{R}$ at a point $x \in D$ is equivalent to Frechet differentiability.

**Proof.** A direct consequence of and theorem 99 and theorem 100.

**Theorem 102.** In $\mathbb{R}^n$, continuous convex functions $f : D \to \mathbb{R}$ are differentiable almost everywhere.

**Proof.** For all $h \in S_X$ and small enough $t > 0$ we define
\[
\varphi_h^t : D \to \mathbb{R} \quad \varphi_h^t(x) := \frac{f(x + th) - f(x)}{t}
\]
\[
\text{and } \varphi_h(x) := f'_+(x)(h) = \lim_{t \to 0} \varphi_h^t(x).
\]

Considered as functions of $x$, $\varphi_h^t$ are obviously continuous hence Borel measurable, and so $\varphi_h$ is also Borel measurable.

Denote by
\[
B_h := \{x \in D : -f'_+(x)(-h) < f'_+(x)(h) = \{x \in D : -\varphi_{-h}(x) - \varphi_h(x) < 0\}
\]
the set of points $x \in D$ where the one-sided derivative $f'_+(x)(h)$ is not linear, given a fixed direction $h \in S_X$. If $B_h$ is nonempty, $f$ is not differentiable at $x$.

The sets $B_h$ are Borel sets since they are the preimages of $(-\infty, 0)$ under a Borel function. We will show that it is a null set for every direction $h$.

Fix $h \in S_X$. Denote by $\delta_x := \sup\{t > 0 : x + th \in D\}$.

The function $t \mapsto f(x + th)$ is a convex function of one variable. By [Phe93, theorem 1.16], it is differentiable $\mu_1$-almost everywhere in $[0, \delta_x)$, where $\mu_m$ is the Lebesgue $m$-measure.

Denote
\[
H := \text{span}\{h\} \cong \mathbb{R}^1,
\]
\[
H^1 \cong \mathbb{R}^{n-1} - \text{the orthogonal complement of } H \text{ in } \mathbb{R}^n,
\]

59
\( L_x := \{ x + th, 0 \leq t < \delta_x \} - \text{half-open segments in } D. \)

The whole domain \( D \) can be represented as \( D = \bigcup \{ L_x : x \in H^1 \}. \)

We can now use Fubini’s theorem to show that \( B_h \) is a null set:

\[
\mu_n(B_h) = \int_{B_h} dz = \int_{\mathbb{R}^n = H^1 \oplus H} \chi_{B_h}(z)d\mathcal{H}^n = \int_{H^1} \int_{L_x} \chi_{B_h}(y) dy dx = \int_{H^1} \mu_1(B_h \cap L_x) dx = \int_{H^1} 0 dx = 0.
\]

Hence, for all \( h \in S_X, -f'(x)(-h) = f'(x)(h) \) for almost all \( x \in D. \)

In particular, if \( e_1, ..., e_n \) is the canonical basis of \( \mathbb{R}^n \), the \( i \)-th partial derivative \( \frac{\partial f}{\partial x_i}(x) \) exists only in \( D B_{e_i}. \)

The gradient

\[
\nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x) \right)
\]

then exists in

\[
\hat{D} := (D B_{e_1}) \cap ... \cap (D B_{e_n}) = D \setminus \left( \bigcup_{i=1}^n B_{e_i} \right).
\]

Corollary 101 then implies that \( f \) is Frechet differentiable in \( \hat{D} \), i.e. almost everywhere in \( D. \) \( \square \)
2.11. Subdifferentials

Let $X$ be a Hausdorff topological vector space, let $D \subseteq X$ be an open set and $f : D \rightarrow \mathbb{R}$ be any function.

**Definition 103.** We fix a point $x \in D$. We define different types of subgradients and subdifferentials. Subgradients are linear functionals $x^* \in X^*$ that approximate $f$ at the point $x$ in a certain way, and a subdifferential is the set of all subgradients of a given type.

**(a)** We say that $x^* \in X^*$ is a subgradient of $f$ at $x$ if for every $y \in D$ we have

$$f(y) - f(x) \geq \langle x^*, y - x \rangle.$$ 

The subdifferential of $f$ at $x$ is denoted by $\partial f(x)$ and is also sometimes called the convex subdifferential because of proposition 96.

**(b)** We say that $x^* \in X^*$ is a Clarke (generalized) subgradient of $f$ at $x$ if for every direction $h \in X$ we have

$$f^*(x)(h) \geq \langle x^*, h \rangle,$$

where $f^*(x)(h)$ is the generalized Clarke derivative.

The subdifferential of $f$ at $x$ is denoted by $\partial_C f(x)$. Confusingly, the Clarke subdifferential is called the “generalized gradient” by Clarke himself with no special name for the Clarke subgradients.

See section 2.12 (Clarke generalized gradients) for properties of these subgradients.

**(c)** We say that $x^* \in X^*$ is a proximal subgradient of $f$ at $x$ if there exist $\sigma > 0$ and a neighborhood $V \subseteq X$ of $x$ such that for every $y \in D \cap V$ we have

$$f(y) - f(x) + \sigma ||y - x||^2 \geq \langle x^*, y - x \rangle.$$ 

The proximal subdifferential of $f$ at $x$ is denoted by $\partial_P f(x)$.

**(d)** Suppose the following are satisfied:

1. $\{x_n\}_n \subseteq D$ is a sequence of points converging to $x$
2. $f(x_n) \rightarrow f(x)$ (redundant if $f$ is continuous)
3. $x^*_n$ is a proximal subgradient for $f$ at $x_n$ for every $n \in \mathbb{Z}_{>0}$.

If the limit $x^* := \lim_n x^*_n$ exists and is a continuous linear functional, we call $x^*$ a limiting subgradient of $f$ at $x$.

The limiting subdifferential of $f$ at $x$ is denoted by $\partial_L f(x)$.
2.12. Clarke generalized gradients

Let $X$ be a Banach space and $f : X \to \mathbb{R}$ be locally Lipschitz.

**Definition 104.** Let $x \in X$ and $U \subseteq X$ be a neighborhood of $x$ where $f$ is $L$-Lipschitz, i.e.

$$\forall y, z \in U, |f(y) - f(z)| \leq L\|y - z\|.$$

We use the Clarke generalized derivative,

$$f^c(x)(h) := \limsup_{t \downarrow 0} \frac{f(y + th) - f(y)}{t}.$$

We define the **generalized gradient** of $f$ at $x$ to be the set

$$\partial C f(x) := \{ x^* \in X^* : \forall h \in X, f^c(x)(h) \geq \langle x^*, h \rangle \}.$$

We say that the vector $h$ is a descent direction of $f$ at $x$ if

$$\limsup_{t \downarrow 0} \frac{f(x + th) - f(x)}{t} < 0.$$

**Proposition 105.** The generalized derivative of a locally Lipschitz function $f : X \to \mathbb{R}$ exists for every $x \in X$.

**Proof.** Let $x, h \in X$ and let $U$ be a neighborhood of $x$ where the Lipschitz condition holds with the constant $L_U$. Then there exists $\delta_0 > 0$ such that $B(x, \delta_0) \subseteq U$.

Define $\delta_1 := \frac{1}{2} \min\left\{ \delta_0, \frac{\delta_0}{\|h\|} \right\} < \delta_0$, so that for $y \in B(x, \delta_1)$ and $t \in (0, \delta_1)$ we have

$$\|y + th - x\| \leq \|y - x\| + t\|h\| \leq \delta_1 + \delta_1\|h\| \leq \begin{cases} \frac{\delta_0}{2} (1 + \|h\|), & \|h\| \leq 1 \\ \frac{\delta_0}{2\|h\|} (1 + \|h\|), & \|h\| > 1. \end{cases}$$

In both cases we get that $y + th \in B(x, \delta_0)$.

The generalized derivative in $x$ in the direction $h \in X$ is then norm-bounded by

$$|f^c(x)(h)| = \limsup_{y \to x} \frac{f(y + th) - f(y)}{t} = \limsup_{\delta \to 0} \sup_{y \in B(x, \delta)} \frac{f(y + th) - f(y)}{t} \leq \sup_{y \in B(x, \delta_1)} \frac{f(y + th) - f(y)}{t} \leq \sup_{y \in B(x, \delta_1)} \frac{\|y + th - (y)\|}{t} = \|h\|.$$

The fact that $f$ is locally Lipschitz gave us that the supremum is taken over a bounded set and thus the generalized derivative exists. \qed
3. Complex analysis

Complex analysis an extension of section 2 (Real analysis), which is concerned with studying functions with values in finite-dimensional complex Hilbert spaces $\mathbb{C}^n$ rather than Euclidean spaces $\mathbb{R}^n$. A lot of results are different, however much of section 2 (Real analysis) is delegated here because it holds in greater generality. For these general results, through the section, $\mathbb{K}$ will refer to either $\mathbb{R}$ or $\mathbb{C}$.

Despite complex analysis being a very rich field, we are mostly concerned with special functions, to which we dedicate the following sections:

- Section 3.3 (Power series)
- Section 3.4 (Trigonometric functions)
- Section 3.5 (Exponential function)
- Section 3.6 (Trigonometric polynomials)
- Section 3.7 (Special functions)
3.1. Complex functions

**Definition 106.** We will define multiple Banach spaces of sequences over $\mathbb{C}$.

(a) The simplest nontrivial sequence space is that of all sequences with only finitely many nonzero elements. It is denoted by $c_{00}$. It can be defined as

$$c_{00} := \bigcup_{i=1}^{\infty} \mathbb{C}^k,$$

where $\mathbb{C}^k$ is the corresponding tuple space. This space can be generalized to modules over semirings.

**Definition 107.** We will define multiple Banach spaces of functions over $\mathbb{K}$.

(a) Define the set of functions **vanishing at infinity**:

$$C_0(\mathbb{C}) := \{f : \mathbb{C} \to \mathbb{C} : f(x) \xrightarrow{|x| \to \infty} 0\}.$$

(b) Fix topological space $X$. The set $C(X) = C(X, \mathbb{K})$ of all $\mathbb{K}$-valued continuous functions on $X$ in a Banach space over $\mathbb{K}$.

**Theorem 108** (Arzela-Ascoli). Let $X$ be a compact Hausdorff space.

A family $\mathcal{F} \subseteq C(X, \mathbb{R})$ of continuous real-valued functions is totally bounded if and only if it is pointwise bounded and equicontinuous.
3.2. Series

Here \((X, ||\cdot||)\) will refer to a Banach space over \(\mathbb{K}\).

**Definition 109.** When extending addition to a countable amount of terms, we need to impose some regularity conditions to avoid contradictions. The topologies of \(\mathbb{R}\) and \(\mathbb{C}\) are complete and allow us to define convergent and divergent series. We define series in great generality because the theory easily allows it.

A **numeric series** or simply **series** is an infinite sequence \(x_0, x_1, \ldots \in X\), which we call **terms**, usually written as

\[
\sum_{k=0}^{\infty} x_k. \tag{29}
\]

To each series, there corresponds its sequence of **partial sums**

\[
S_n := \sum_{k=0}^{n} x_k, n = 0, 1, 2, \ldots
\]

We can equivalently define a series as a sequence of partial sums and then recover the terms as

\[
x_k := \begin{cases} 
S_0, & k = 0, \\
S_k - S_{k-1}, & k > 0
\end{cases}
\]

We say that the series (29) **converges** to a value \(x\) if \(\lim_{n \to \infty} S_n = x\) in the sense of **definition** 274 (b). The value \(x\) is called the **sum** of the series.

If a series does not converge, we say that it is **divergent**. If the related series

\[
\sum_{k=0}^{\infty} ||x_k|| \tag{30}
\]

converges, we say that (29) is **absolutely convergent**.

**Example 110.** Several examples of series are

- An absolutely convergent series is (537).
- A divergent series is the harmonic series (40).
- A convergent, but not absolutely convergent series is the alternating harmonic series (41).

**Proposition 111.** An absolutely convergent series is convergent.

**Proof.** Suppose that (30) converges.

By the triangle inequality, for each index \(n\) we have

\[
||\sum_{k=0}^{n} x_k|| \leq \sum_{k=0}^{n} ||x_k|| \leq \sum_{k=0}^{\infty} ||x_k||.
\]
Thus, the sequence \( \left\{ \| \sum_{k=0}^{n} x_k \| \right\}_{n=0}^{\infty} \) is a bounded (by \( \sum_{k=0}^{\infty} \| x_k \| \)) monotone sequence, which by proposition 56 is convergent.

Therefore, the series (29) is convergent.

\[ \square \]

**Remark 112.** Convergence of the series (29) can be established using the convergence of the nonnegative series (30).

The convergence of the latter can be established using techniques in section 2.5 (Real series) like proposition 65 (Cauchy’s root test) or proposition 66 (d’Alambert’s ratio test).

**Proposition 113.** For every series (29) we have

\[
\| \sum_{k=0}^{\infty} x_k \| \leq \sum_{k=0}^{\infty} \| x_k \|,
\tag{31}
\]

where both limits are allowed to be infinite.

**Proof.** If the series on the right diverges, the inequality is obviously true.

Suppose that it is convergent. By proposition 111, the limit (29) exists.

By the triangle inequality, for each index \( n \) we have

\[
\| \sum_{k=0}^{n} x_k \| \leq \sum_{k=0}^{n} \| x_k \|.
\]

By proposition 54, since both sequences are convergent, we obtain eq. (31).

\[ \square \]

**Proposition 114.** The terms of the convergent series (29) vanish as \( k \to \infty \), that is,

\[
\lim_{k \to \infty} x_k = 0.
\]

**Proof.** Since the series is convergent, its sequence of partial sums converges, i.e. the partial sums get arbitrarily close to each other. Then

\[
\| x_n \| = \| S_n - S_{n-1} \| \to 0.
\]

\[ \square \]

**Theorem 115.** Consider two convergent series

\[
A := \sum_{k=0}^{\infty} x_k \tag{32}
\]

and

\[
B := \sum_{k=0}^{\infty} y_k. \tag{33}
\]

If either eq. (32) or eq. (33) converges absolutely, then

\[
\sum_{k=0}^{\infty} \sum_{m=0}^{k} x_m y_{k-m} = AB. \tag{34}
\]

66
Proposition 116 (Cauchy’s series convergence criterion). The series (29) converges if and only if for every $\varepsilon > 0$ there exists an index $k_0$ such that

$$\| \sum_{k=m}^{n} x_k \| < \varepsilon \quad \forall m, n \geq k_0.$$  

Proof. This is simply a restatement of theorem 353 (Cauchy’s net convergence criterion). \qed

Proposition 117 (Cauchy’s series continuity criterion). Fix a topological space $A$ and a set $S \subset A$. Let $\{ f_k \}_{k=0}^{\infty}$ be a sequence of continuous functions from $S$ to $X$.

Define the function $f : S \to X$ as

$$f(x) := \sum_{k=0}^{\infty} f_k(x). \quad (35)$$

A sufficient condition for $f$ to be continuous in $S$ is that for every $\varepsilon > 0$ there exists an index $K$ such that

$$\| \sum_{k=m}^{n} f_k(x) \| < \varepsilon \quad \forall m, n \geq K$$

simultaneously for all $x \in S$.

Proof. This is simply a restatement of proposition 347 in the style of proposition 116 (Cauchy’s series convergence criterion). \qed

Corollary 118 (Weierstrass’ series criterion). Let $S$ be any set and $\{ f_k \}_{k=0}^{\infty}$ be a sequence of functions from $S$ to $X$. Consider the series eq. (35). If

$$\forall k \in \mathbb{Z}^{>0} \exists M_k \in \mathbb{R}^{>0} \forall x \in S : \| f_k(x) \| < M_k$$

and if the series

$$\sum_{k=0}^{\infty} M_k \quad (36)$$

converges, then the limit eq. (35) exists for every $x \in S$ and, furthermore, the series converges absolutely and uniformly.

In analogy to proposition 64, we say that the series eq. (36) dominates the series eq. (35).

In particular, if $S$ has a topology and the functions $f_k(x), k = 0, 1, \ldots$ are continuous (resp. uniformly continuous), so is $f(x)$.

Proof. By proposition 64, the series

$$\sum_{k=0}^{\infty} \| f_k(x) \|$$

converges for any $x \in S$, hence eq. (35) converges absolutely for any $x \in S$.

Furthermore, each of the functions $f_k(x)$ is bounded by $B(0, M_k)$ and $M_k$ does not depend on $x$, hence the convergence is uniform.

The rest of the theorem follows from proposition 347. \qed
Corollary 119. Let \( X \subseteq \mathbb{R} \) be a nonempty set. Consider the series of real-valued real functions
\[
f(x) := \sum_{k=0}^{\infty} f_k(x) / 2^k,
\]
where \( \{f_k\}_{k=0}^{\infty} \subseteq B_C(X) \) is a sequence of continuous functions bounded in \([-1, 1]\).
Then \( f(x) \) is defined and continuous for all \( x \in X \).

Proof. For \(|x| \leq 1\), the series is dominated by the geometric series eq. (538), which sums to 2, hence by corollary 118 (Weierstrass’ series criterion) \( f(x) \) is continuous in the interval \([-1, 1]\).
Note that
\[
f(2x) := \sum_{k=0}^{\infty} x / 2^{k-1} = 2f(x),
\]
hence the series eq. (37) also converges for \(|x| \leq 2\).
By induction on \( n \), we show that \( f(2^n x) = 2^n f(x) \) and thus \( f(x) \) is continuous in \( B(0, 2^n) \), therefore also on the entire real line \( \mathbb{R} \).

Example 120. Consider the real series
\[
f(x) := \sum_{k=0}^{\infty} x^k(1 - x).
\]
It converges for \(|x| < 1\) because it is dominated by a convergent geometric series.
For \( x \in (0, 1) \),
\[
f(x) = \sum_{k=0}^{\infty} x^k(1 - x) = \sum_{k=0}^{\infty} x^k - \sum_{k=1}^{\infty} x^k = 1.
\]
But
\[
limit_{t \to 1} f(x) = 1 \neq 0 = f(1) = f(limit \, t).
\]
This shows that \( f(x) \) is not continuous, despite every term being continuous.
By contraposition to corollary 118 (Weierstrass’ series criterion), it follows that no series
that dominates \( f(x) \) converges.

Theorem 121. Fix a uniform space \((A, \mathcal{U})\) and let \( S \subseteq A \). Let \( f_k : S \to X, k = 0, 1, \ldots \) be a sequence of functions and assume that \( x_0 \in M \) is a limit point of each of these functions.

(a) If the sequence \( \{f_k\}_{k=0}^{\infty} \) converges uniformly on \( S \), we can exchange the limits
\[
\lim_{x \to x_0} \lim_{k \to \infty} f_k(x) = \lim_{k \to \infty} \lim_{x \to x_0} f_k(x).
\]

(b) If the series eq. (35) converges uniformly on \( S \), we can exchange the limits
\[
\lim_{x \to x_0} \sum_{k=0}^{\infty} f_k(x) = \sum_{k=0}^{\infty} \lim_{x \to x_0} f_k(x).
\]
Remark 122. If the functions $f_k$ in theorem 121 (b) are continuous at $x_0$, we have the additional equality
\[
\lim_{x \to x_0} f(x) = \lim_{x \to x_0} \sum_{k=0}^{\infty} f_k(x) = \sum_{k=0}^{\infty} \lim_{x \to x_0} f_k \left( \lim_{x \to x_0} x \right) = f \left( \lim_{x \to x_0} x \right),
\]
thus $f$ is continuous at $x_0$. The continuity actually follows from proposition 117 (Cauchy’s series continuity criterion) directly.

Corollary 123. Let $\{ f_k \}_{k=0}^{\infty} \subseteq C([a, b], \mathbb{R})$.

(a) If the sequence $\{ f_k \}_{k=0}^{\infty}$ converges uniformly, then
\[
\lim_{k \to \infty} \int_{a}^{b} f_k(x) \, dx = \int_{a}^{b} \lim_{k \to \infty} f_k(x) \, dx.
\]

(b) If the series eq. (35) converges uniformly, then
\[
\int_{a}^{b} f(x) \, dx = \int_{a}^{b} \sum_{k=0}^{\infty} f_k(x) \, dx = \sum_{k=0}^{\infty} \int_{a}^{b} f_k(x) \, dx.
\]

Proof.

Proof of 123 (a). Assume that the sequence $\{ f_k \}_{k=0}^{\infty}$ converges uniformly to $f$. Then by theorem 121, we note that for any index $k$, the difference $r_k(x) := f(x) - f_k(x)$ is continuous, hence integrable, and
\[
\int_{a}^{b} f(x) \, dx = \int_{a}^{b} f_k(x) \, dx + \int_{a}^{b} r_k(x) \, dx.
\]
Because of the uniform convergence, for any $\delta > 0$ and there exist an index $k_0$ such that
\[
|f(x) - f_k(x)| = |r_k(x)| < \delta \quad \forall k \geq k_0, \forall x \in [a, b].
\]
Then
\[
|\int_{a}^{b} f(x) \, dx - \int_{a}^{b} f_k(x) \, dx| = |\int_{a}^{b} r_k(x) \, dx| < (b - a)\delta.
\]
Given $\varepsilon > 0$, we define $\delta := \frac{\varepsilon}{b - a}$ to obtain an index $k_0$ such that
\[
|\int_{a}^{b} f(x) \, dx - \int_{a}^{b} f_k(x) \, dx| = |\int_{a}^{b} r_k(x) \, dx| < \varepsilon \quad \forall k \geq k_0.
\]
Thus, definition 274 (b) is satisfied and equality holds.

Proof of 123 (b). This is a special case of corollary 123 (a).
Corollary 124. Let \( \{ f_k \}_{k=0}^{\infty} \subseteq C^1([a,b], \mathbb{R}) \). Suppose that the series eq. (35) converges for at least one point \( x_0 \in [a,b] \).

(a) If the sequence \( \{ Df_k \}_{k=0}^{\infty} \) of derivatives converges uniformly, then \( \{ f_k \}_{k=0}^{\infty} \) also converges uniformly, its limit is differentiable in \( (a,b) \) and

\[
D \left( \lim_{k \to \infty} f_k(x) \right) = \lim_{k \to \infty} Df_k(x).
\]

(b) If the series of derivatives

\[
\sum_{k=0}^{\infty} Df_k(x)
\]

converges uniformly, then eq. (35) converges uniformly, is differentiable in \( (a,b) \) and

\[
D \left( \sum_{k=0}^{\infty} f_k(x) \right) = \sum_{k=0}^{\infty} Df_k(x).
\]

Proof.

Proof of 124 (a). Fix \( \varepsilon > 0 \). Since the sequence \( \{ f_k \}_{k=0}^{\infty} \) converges for \( x_0 \), there exists an index \( k_0 \) such that

\[
|f_m(x_0) - f_n(x_0)| < \varepsilon \quad \forall m, n \geq k_0.
\]

Furthermore, there exists an index \( k_1 \) such that

\[
|Df_m(x) - Df_n(x)| < \varepsilon \quad \forall x \in [a,b] \forall m, n \geq k_0.
\]

Fix \( m, n \geq k_0 \) and \( x \in [a,b] \). Note that the function \( f_m - f_n \) is differentiable and thus by the mean value theorem, there exists \( \xi \) between \( x_0 \) and \( x \) such that

\[
\frac{|f_m(x) - f_n(x)| - [f_m(x_0) - f_n(x_0)]}{x - x_0} = Df_m(\xi) - Df_n(\xi).
\]

Thus,

\[
|f_m(x) - f_n(x)| \leq |f_m(x_0) - f_n(x_0)| + (x - x_0)|Df_m(\xi) - Df_n(\xi)| < 2(x - x_0)\varepsilon \leq 2(b - a)\varepsilon.
\]

Therefore, the limit \( \lim_{k \to \infty} f_k(x) \) exists. Since \( x \) was arbitrary and \( 2(b - a)\varepsilon \) does not depend on \( x \), we conclude that

\[
f(x) := \lim_{k \to \infty} f_k(x)
\]

is uniformly convergent on \([a,b]\).
By the Newton-Leibniz theorem, for the sequence \( \{Df_k\}_{k=0}^{\infty} \) of derivatives we have
\[
\lim_{k \to \infty} \int_a^x Df_k(t) \, dt = \lim_{k \to \infty} [f_k(x) - f_k(a)] = \lim_{k \to \infty} f_k(x) - \lim_{k \to \infty} f_k(a) = f(x) - f(a).
\]
Differentiating both sides, we obtain
\[
D \left( \lim_{k \to \infty} \int_a^x Df_k(t) \, dt \right) = D \left( \lim_{k \to \infty} f_k(x) \, dt \right) = Df(x).
\]
Corollary 123 (a) allows us to conclude that
\[
Df(x) = D \left( \lim_{k \to \infty} \int_a^x Df_k(t) \, dt \right) = D \int_a^x \lim_{k \to \infty} Df_k(t) \, dt = \lim_{k \to \infty} Df_k(x).
\]

**Proof of 124 (b).** This is a special case of corollary 123 (a). \(\square\)

**Example 125.** We list several important series related to harmonic progressions.

(a) The series
\[
\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \tag{40}
\]
is called the **harmonic series**. It diverges as shown in proposition 126, which make it much less useful in practice, however it is an important enough example that it has a dedicated name.

(b) The series
\[
\sum_{k=1}^{\infty} \frac{(-1)^k}{k} = \sum_{m=1}^{\infty} \left( \frac{1}{2m-1} - \frac{1}{2m} \right) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \tag{41}
\]
is called the **alternating harmonic series**. It converges, but not absolutely — proposition 127.

(c) For any \( s \in \mathbb{C} \), the series
\[
\sum_{k=1}^{\infty} \frac{1}{k^s}.
\]
is called the **hyperharmonic series**.

Unlike the harmonic series, the hyperharmonic series sometimes converges — see proposition 128.

**Proposition 126.** The harmonic series (40) diverges.
Proof. Define the series

\[
1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{16} + \cdots + \frac{1}{16}\right) + \cdots
\]

It is divergent as the sum of infinitely many \(\frac{1}{2}\). Furthermore, it is dominated by the harmonic series:

\[
1 + \frac{1}{2} + \frac{1}{3} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6}\right) + \left(\frac{1}{7} + \frac{1}{8}\right) + \left(\frac{1}{9} + \cdots + \frac{1}{16}\right) + \cdots
\]

\[
1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \cdots + \frac{1}{16} + \cdots
\]

Thus, by Proposition 64, the harmonic series also diverges.

Proposition 127. \([\text{Fic68b, 6247}]\) Consider the alternating harmonic series (41).
Compare the series with (40). Note that, by Proposition 69 (Leibniz’ alternating series test), the series is convergent. It is not absolutely convergent, because the harmonic series (40) is divergent.

Proposition 128. The hyperharmonic series (125 (c)) converges all \(s \in \mathbb{C}\) with \(\text{real}(s) > 1\).

Proof. Let \(s = (1 + \varepsilon) + bi\). We use the integral test on the series \(\sum_{k=1}^{\infty} |k|^{-s}\):

\[
\int_{1}^{\infty} \frac{1}{|x|^{s}} \, dx = \int_{1}^{\infty} \frac{1}{x^{1+\varepsilon} |x|^{b} |x|} \, dx = -\frac{1}{\varepsilon x^{\varepsilon}} \bigg|_{x=1}^{\infty} = \frac{1}{\varepsilon} \lim_{x \to \infty} \left(1 - \frac{1}{x^{\varepsilon}}\right) = \frac{1}{\varepsilon}.
\]

The integral is finite, hence the hyperharmonic series is absolutely convergent.
3.3. Power series

**Definition 129.** Let $\mathbb{K}[X]$ be the space of formal power series defined in definition 578. To each formal power series

$$\sum_{k=0}^{\infty} a_k X^k$$

there corresponds a function, called a **power series**

$$p(x) := \sum_{k=0}^{\infty} a_k x^k.$$ \hspace{1cm} (43)

We sometimes slightly generalize this notion slightly by using a “shift” by $\alpha \in \mathbb{K}$: define the function

$$p(x) := \sum_{k=0}^{\infty} a_k (x - \alpha)^k.$$ \hspace{1cm} (44)

If the limit exists (as a numeric series) for a certain $x \in \mathbb{K}$, we say that the series **converges** at $x$.

The series is no longer “formal” because it is now a proper function instead of an abstract algebraic object, although a power series may only be defined in a subset of $\mathbb{K}$ (that is, a partial function).

**Theorem 130.** For every power series (43), there exists a nonnegative extended real number $r \in [0, +\infty]$, called its **radius of convergence**, such that (43) converges absolutely if $|x| < r$ and diverges if $|x| > r$.

The behavior of the series is more complicated when $|x| = r$ (unless $r = 0$, in which case the power series converges if and only if $x = 0$).

**Proof.** Define

$$q := \limsup_{n \to \infty} \sqrt[n]{|a_n|},$$

where we put $q = +\infty$ if the limit does not exist. We have

$$\limsup_{n \to \infty} \sqrt[n]{|x^n a_n|} = |x|^q.$$

By proposition 65 (Cauchy’s root test), (43) converges absolutely if $|z|^q < 1$ and diverges if $|z|^q > 1$.

Thus, $r := \frac{1}{q}$ is the desired radius of convergence.

Note that we may also use proposition 66 (d’Alambert’s ratio test) for finding the same radius of convergence by proposition 67.

**Proposition 131.** Power series of the form

$$f_o(z) := \sum_{m \text{ is odd}} a_m z^m = \sum_{k=0}^{\infty} a_{2k+1} z^{2k+1}.$$ \hspace{1cm} (45)
are odd functions and power series of the form
\[ f_e(z) := \sum_{m \text{ is even}} a_m z^m = \sum_{k=0}^{\infty} a_{2k} z^{2k} \quad (46) \]

are even functions.

**Proof.** If (45) converges for \( z \in \mathbb{C} \),
\[ f_o(-z) = \sum_{k=0}^{\infty} a_{2k+1} (-z)^{2k+1} = \sum_{k=0}^{\infty} a_{2k+1} (-1)^{2k+1} z^{2k+1} = -\sum_{k=0}^{\infty} a_{2k+1} z^{2k+1} = -f_o(z). \]

Analogously, since \((-1)^{2k} = 1\), we have \( f_e(-z) = f_e(z) \). \( \square \)

**Proposition 132.** A power series is locally uniformly convergent in the interior of its domain of convergence.

**Proof.** Assume that the series (43) converges inside the ball \( B(0, R) \). Fix \( x \in B(0, R) \) and \( R_x < R - |x| \). Then the geometric series
\[ \sum_{k=0}^{\infty} a_k R_x^k \]
converges and dominates (43) in the ball \( B(x, R_x) \). Thus, by corollary 118 (Weierstrass’ series criterion), (43) converges uniformly in \( B(x, R_x) \).

Since the choice of \( x \in B(0, R) \) was arbitrary, we conclude that (43) is locally uniformly convergent. \( \square \)

**Theorem 133.** Suppose that the power series (43) has a (potentially infinite) radius of convergence \( R \).

(a) \( p(x) \) is differentiable in \( B(0, R) \) and can be differentiated termwise as
\[ p'(x) = \sum_{k=0}^{\infty} a_{k+1} (k+1)x^k. \quad (47) \]

Furthermore, \( p'(x) \) has the same radius of convergence as \( p(x) \).

(b) If the series is real and \( |x| < R \), \( p(x) \) is integrable in \([0, x]\) (or \([x, 0]\)) and can be integrated termwise as
\[ \int_0^x p(t)dt = \sum_{k=0}^{\infty} a_k \frac{x^{k+1}}{k+1}. \quad (48) \]

**Proof.**
**Proof of 133 (a).** Note that the right-hand side of eq. (47) is a power series. Furthermore, its radius of convergence is, by theorem 130,

\[
\lim_{k \to \infty} \left| \frac{a_{k+1}(k + 1)x^k}{a_{k+2}(k + 2)x^{k+1}} \right| = |x| \lim_{k \to \infty} \frac{k + 1}{k + 2} \left| \frac{a_{k+1}}{a_{k+2}} \right| = R.
\]

Fix \( x \in B(0,R) \) and choose \( r \in (|x|,R) \). Both series are uniformly convergent in \( B(0,r) \). By corollary 124 (a), the equality eq. (47) holds in \( B(0,r) \), hence it also holds for \( x \).

**Proof of 133 (b).** Analogously to theorem 133 (a), we conclude that the right-hand side of eq. (48) is a power series with radius of convergence \( R \).

The rest follows directly from corollary 123. \( \square \)
3.4. Trigonometric functions

**Definition 134.** We define the two basic **trigonometric functions**. They are also called **circular trigonometric functions** to distinguish them from the hyperbolic trigonometric functions defined and motivated in definition 140.

(a) The **sine** function, also called the **sinus** function, is

\[
\sin(z) := -i \sum_{m \text{ is odd}} \frac{i^m z^m}{m!} = -i \sum_{k=0}^{\infty} \frac{i^{2k+1} z^{2k+1}}{(2k+1)!} = \sum_{k=0}^{\infty} \frac{i^{2k} z^{2k+1}}{(2k+1)!}
\]

(b) The **cosine** function, also called the **cosinus** function, is

\[
\cos(z) := \sum_{m \text{ is even}} \frac{i^m z^m}{m!} = \sum_{k=0}^{\infty} \frac{i^{2k} z^{2k}}{(2k)!}
\]

Definition 430 justifies the term “angle” for the parameter of the trigonometric functions.

**Proposition 135.** The **main trigonometric functions** have the following basic properties:

(a) Both \( \sin(z) \) and \( \cos(z) \) converge in the entire complex plane.

(b) \( \sin(z) \) is an odd function and \( \cos(z) \) is an even function.

(c) \( \sin'(z) = \cos(z) \) and \( \cos'(z) = -\sin(z) \) for all \( z \in \mathbb{C} \).

**Proof.**

**Proof of 135 (a).** Note that the zero coefficients in the expansion of either \( \sin \) or \( \cos \) do not alter convergence. Therefore, by theorem 130, the radius of convergence is

\[
\limsup_{k \to \infty} \frac{|i^{k-1} k!|}{|i^k (k-1)!|} = \limsup_{k \to \infty} k = +\infty.
\]

**Proof of 135 (b).** Follows from proposition 131.

**Proof of 135 (c).** Follows from proposition 132 and corollary 124. \( \square \)

**Proposition 136.** We have the following basic trigonometric identities:

(a) (Pythagorean identity) For any \( z \in \mathbb{C} \),

\[
\sin(z)^2 + \cos(z)^2 = 1. \quad (49)
\]

(b) (Products) For \( x, y \in \mathbb{C} \),

\[
2 \sin(x) \sin(y) = \cos(x - y) - \cos(x + y) \quad (50)
\]

\[
2 \cos(x) \cos(y) = \cos(x - y) + \cos(x + y) \quad (51)
\]

\[
2 \sin(x) \cos(y) = \sin(x - y) + \sin(x + y) \quad (52)
\]

\[
2 \cos(x) \sin(y) = -\sin(x - y) + \sin(x + y) \quad (53)
\]

76
\( (c) \) **Sums** For \( x, y \in \mathbb{C} \),

\[
\begin{align*}
\sin(x) + \sin(y) &= 2 \cos\left(\frac{x-y}{2}\right) \sin\left(\frac{x+y}{2}\right) \\
\sin(x) - \sin(y) &= 2 \sin\left(\frac{x-y}{2}\right) \cos\left(\frac{x+y}{2}\right) \\
\cos(x) + \cos(y) &= 2 \cos\left(\frac{x-y}{2}\right) \cos\left(\frac{x+y}{2}\right) \\
\cos(x) - \cos(y) &= -2 \sin\left(\frac{x-y}{2}\right) \sin\left(\frac{x+y}{2}\right)
\end{align*}
\] (54) (55) (56) (57)

\( (d) \) **Sum of angles** For \( x, y \in \mathbb{C} \),

\[
\begin{align*}
\sin(x+y) &= \cos(x) \sin(y) + \cos(y) \sin(x) \\
\cos(x+y) &= \cos(x) \cos(y) - \sin(x) \sin(y)
\end{align*}
\] (58) (59)

**Proof.** We first use Cauchy multiplication for the power series \( \cos(u) \) and \( \cos(w) \):

\[
\cos(u) \cos(w) = \left( \sum_{k=0}^{\infty} \frac{i^{2k} u^{2k}}{(2k)!} \right) \left( \sum_{k=0}^{\infty} \frac{i^{2k} w^{2k}}{(2k)!} \right) =
\]

\[
= \sum_{k=0}^{\infty} \sum_{m=0}^{k} \frac{i^{2m} u^{2m} i^{2(k-m)} w^{2(k-m)}}{(2m)! (2(k-m))!} =
\]

\[
= \sum_{k=0}^{\infty} \frac{i^{2k}}{(2k)!} \sum_{m=0}^{k} \left( \frac{2k}{2m} \right) u^{2m} w^{2(k-m)}. \quad (60)
\]

Analogously,

\[
\sin(u) \sin(w) = (-i)(-i) \left( \sum_{k=0}^{\infty} \frac{i^{2k+1} u^{2k+1}}{(2k+1)!} \right) \left( \sum_{k=0}^{\infty} \frac{i^{2k+1} w^{2k+1}}{(2k+1)!} \right) =
\]

\[
= -\sum_{k=0}^{\infty} \sum_{m=0}^{k} \frac{i^{2m+1} u^{2m+1} i^{2(k-m)+1} w^{2(k-m)+1}}{(2m+1)! (2(k-m)+1)!} =
\]

\[
= -\sum_{k=0}^{\infty} \frac{i^{2(k+1)}}{(2(k+1))!} \sum_{m=0}^{k} \frac{2(k+1)}{2m+1} u^{2m+1} w^{2(k-m)+1} =
\]

\[
= -\sum_{k=1}^{\infty} \frac{i^{2k}}{(2k)!} \sum_{m=0}^{k-1} \left( \frac{2k}{2m+1} \right) u^{2m+1} w^{2(k-m)}. \quad (61)
\]

**Proof of 136 (a).** From (60) and (61) we have

\[
\sin(z)^2 + \cos(z)^2 = 1 + \sum_{k=1}^{\infty} \frac{i^{2k} z^{2k}}{(2k)!} \left[ -\sum_{m=0}^{k-1} \frac{2k}{2m+1} + \sum_{m=0}^{k} \frac{2k}{2m} \right].
\]

\[= a_k \]
It remains to show that the expression $a_k$ equals zero for all $k = 1, 2, \ldots$. We have
\[
a_k = \sum_{m=0}^{k} \binom{2k}{2m} - \sum_{m=0}^{k-1} \binom{2k}{2m+1} = \sum_{m=0}^{k} (-1)^m \binom{2k}{m} = 1 - 1 = 0.
\]

Equation (49) follows.

**Proof of 136 (b).** We will only prove (51) because the other identities are proved analogously. We have
\[
\cos(v - w) + \cos(v + w) = \sum_{k=0}^{\infty} \frac{i^{2k}}{(2k)!} \left[(v - w)^{2k} + (v + w)^{2k}\right] = \\
= \sum_{k=0}^{\infty} \frac{i^{2k}}{(2k)!} \sum_{m=0}^{2k} \binom{2k}{m} u^{2k-m} w^m \left[(-1)^m + 1\right] = \\
= 2 \sum_{k=0}^{\infty} \frac{i^{2k}}{(2k)!} \sum_{m=0}^{2k} \binom{2k}{m} u^{2k-m} w^m = 2 \cos(v) \cos(w).
\]

**Proof of 136 (c).** Fix some $v, w \in \mathbb{C}$ and define
\[
x := \frac{v + w}{2}, \quad y := \frac{v - w}{2}
\]
so that $v = x + y$ and $w = x - y$.

The identity (54) follows from (52) applied to $x$ and $y$. The other identities are proved analogously.

**Proof of 136 (d).** We will only prove (58) because (59) is proved analogously. From (53),
\[
\sin(x + y) = 2 \cos(x) \sin(y) + \sin(x - y) = 2 \cos(x) \sin(y) + 2 \cos(x) \sin(y) - \sin(x + y).
\]

After dividing by 2, we obtain (58).

**Lemma 137.** We have the following important special values:
\[
\sin(0) = 0, \quad \cos(0) = 1, \quad \sin(\pi) = 0, \quad \cos(\pi) = -1.
\]

**Proof.** (62) follows directly from definition 134. Now consider the restriction of $\cos$ to the real line. Since $\cos(0) \neq 0$ and $\cos$ is continuously differentiable as a power series, in some neighborhood $U$ of 0 we have $0 \not\in \cos(U)$. Therefore, the inverse function theorem holds and there exists a neighborhood $V \subseteq U$ of 1 such that the continuously differentiable function $f : V \to \mathbb{R}$ is the inverse of $\cos$ in $V$ (we have not yet defined $\arccos$). If $y = \cos(x)$, then
\[
Df(y) = \frac{1}{D \cos(x)} = \frac{1}{1 - \sin(x)} = -\frac{1}{\sqrt{1 - y^2}}, \quad y \in \cos(V).
\]

78
The derivative is actually well-defined and continuous anywhere except for \( y \in \{-1, 1\} \).

Therefore, for any \( \alpha \in (-1, 1) \),

\[
f(y) = f(\alpha) - \int_{\alpha}^{y} \frac{1}{\sqrt{1 - t^2}} \, dt, \quad y \in [\alpha, 1).
\]

We already know that \( \cos(0) = 1 \), hence \( f(1) = 0 \) and, since \( f(y) \) is given by a convergent integral in \([\alpha, 1)\), we can extend this interval to \([\alpha, 1]\).

By taking \( y = \alpha \), we obtain

\[
f(y) - f(-y) = -\int_{-y}^{y} \frac{1}{\sqrt{1 - t^2}} \, dt, \quad y \in [-1, 1].
\]

Note that by our definition of \( \pi \),

\[
\pi = \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} \, dt = -[f(1) - f(-1)] = f(-1).
\]

Hence, \( \cos(\pi) = -1 \). From proposition 136 (a),

\[
|\sin(\pi)| = \sqrt{1 - \cos(\pi)^2} = 0,
\]

proving that \( \sin(\pi) = 0 \).

This concludes the proof of (63).

\[\square\]

**Definition 138.** A function \( f : G \to H \) between abelian groups is called **periodic** with period \( p \in G \) if, for all \( x \in G \), we have \( f(x) = f(x + r) \).

The **base period** of a function is the least of all periods, if a minimum exists. When referring to “the period”, we mean the base period.

We can define periods for arbitrary magmas rather than abelian groups, but the definition would make it difficult to talk about the base period.

**Theorem 139.** Both \( \sin(z) \) and \( \cos(z) \) are \( 2\pi \)-periodic.

**Proof.** We will temporarily restrict ourselves to the real line. Since \( \cos(x) \) is continuous, \( \cos^{-1}(\{0\}) \) is a closed set by theorem 322 (Weierstrass’ extreme value theorem) there exists a minimum \( \gamma \) of \([0, \pi] \cap \cos^{-1}(\{0\})\).

Since, by lemma 137, \( \cos(0) = 1 \), it follows that \( \cos(x) > 0, x \in (-\gamma, \gamma) \). Therefore, its primitive function \( \sin(x) \) increases on the same interval. It is also continuous, hence by proposition 136 (a), \( \sin(\gamma) = 1 \) because \( \cos(\gamma) = 0 \).

Because \( \sin \) is an odd function, \( \sin(-\gamma) = \sin(\gamma) = -1 \).

From our definition of \( \pi \) it follows that

\[
\pi = \int_{-1}^{1} \frac{1}{1 - t^2} \, dt = \int_{-\gamma}^{\gamma} \frac{\cos(\varphi)}{1 - \sin(\varphi)^2} \, d\varphi \overset{136 (a)}{=} \int_{-\gamma}^{\gamma} d\varphi = 2\gamma.
\]

79
In order for a number $p$ to be a period of $\sin$, we need to have $\sin(p) = \sin(0) = 0$. But we showed that $\sin(x)$ is increasing from 0 to $\frac{\pi}{2}$ and cannot possibly contain zeros in that interval. Hence, $p > \frac{\pi}{2}$.

We also have $\cos\left(\frac{\pi}{2}\right) = 0$. By proposition 136 (c),

$$\sin\left(\frac{\pi}{2} + x\right) = \sin\left(\frac{\pi}{2}\right) \cos(x) + \cos\left(\frac{\pi}{2}\right) \sin(x) = \cos(x).$$

Since $\cos$ is positive on $[0, \frac{\pi}{2})$, $\sin$ is positive on $[\frac{\pi}{2}, \pi)$.

We already showed in lemma 137 that $\sin(\pi) = 0$.

It follows that the minimal period of $\sin$ is either $\pi$ or a multiple of $\pi$. It cannot be $\pi$ since $\cos(\pi) \neq \cos(0)$, therefore it must be $2\pi$.

**Definition 140.** In analogy with proposition 150, we define **hyperbolic trigonometric functions**.

(a) The **hyperbolic sine** function:

$$\sinh(x) := \frac{e^x - e^{-x}}{2}$$

(b) The **hyperbolic cosine** function:

$$\cosh(x) := \frac{e^x + e^{-x}}{2}$$

Compare eq. (145) and eq. (147) for a justification of the naming.

**Definition 141.** In addition to $\sin(z)$ and $\cos(z)$, we define two additional functions, also called “trigonometric”.

(a) The **partial tangent** function, also called **tangens**, is

$$\tan(z) := \frac{\sin(z)}{\cos(z)}.$$  

It is defined in $\mathbb{C} \setminus (\frac{\pi}{2} + \pi \mathbb{Z})$.

(b) The **partial cotangent function**, also called **cotangens**, is

$$\cot(z) := \frac{\cos(z)}{\sin(z)}.$$  

It is defined in $\mathbb{C} \setminus \pi \mathbb{Z}$.

**Definition 142.** We can define **inverse trigonometric functions**. We will thus restrict ourselves only to real numbers. Fix an integer $k$. Unless noted otherwise, we assume $k = 0$.  

---

80
(a) The **arcus sinus** function $\text{arcsin}(x)$ is defined as the inverse function of $\sin(x)$ (see definition 134 (a)) from $[-1, 1]$ to $\left(\left(k - \frac{1}{2}\right)\pi, \left(k + \frac{1}{2}\right)\pi\right)$.

(b) The **arcus cosinus** function $\text{arccos}(x)$ is defined as the inverse of $\cos(x)$ (see definition 134 (b)) from $[-1, 1]$ to $(k\pi, (k + 1)\pi)$.

(c) The **arcus tangens** function $\text{arctan}(x)$ is defined as the inverse of $\tan(x)$ (see definition 141 (a)) from $\mathbb{R}$ to $\left((k - \frac{1}{2})\pi, (k + \frac{1}{2})\pi\right)$.

(d) The **arcus cotangens** function $\text{arccot}(x)$ is defined as the inverse of $\cot(x)$ (see definition 141 (b)) from $\mathbb{R}$ to $(k\pi, (k + 1)\pi)$.

(e) The **two-argument arcus tangens** function $\text{arctan2}(y, x)$ is a bit special, however it is very useful in practice - see proposition 143. It is defined as

$$\text{arctan2} : \mathbb{R}^2 \setminus \{0\} \to [2k\pi, 2k\pi + 2)$$

$$\text{arctan2}(y, x) := \begin{cases} 
\begin{align*}
\arctan\left(\frac{y}{x}\right), & x > 0 \\
\arctan\left(\frac{y}{x}\right) + \pi, & x < 0 \text{ and } y \geq 0 \\
\arctan\left(\frac{y}{x}\right) - \pi, & x < 0 \text{ and } y < 0 \\
\pi, & x = 0 \text{ and } y \geq 0 \\
-\pi, & x = 0 \text{ and } y < 0.
\end{align*}
\end{cases}$$

**Proposition 143.** Fix an integer $k$. Given $(x_0, y_0) \in S_{\mathbb{R}^2}$, $t_0 := \text{arctan2}(y_0, x_0)$ is the unique, solution to the equation

$$\begin{cases} 
x_0 = \cos(t) \\
y_0 = \sin(t)
\end{cases}$$

in $t \in [2k\pi, 2k\pi + 2)$.
3.5. Exponential function

**Definition 144.** We define the exponential function

\[
\exp(z) := \sum_{k=0}^{\infty} \frac{z^k}{k!}
\]  

(65)

and Euler’s number

\[ e := \exp(1) = \sum_{i=0}^{k} \frac{1}{k!}. \]

Proposition 145 (d) justifies the notation \( e^z = \exp(z) \).

**Proof.** We will show that \( \exp(z) \) converges everywhere. By theorem 130, the radius of convergence is

\[
\limsup_{k \to \infty} k! = \limsup_{k \to \infty} k = +\infty
\]

Hence, the radius of convergence of \( \exp(x) \) is infinite. \( \Box \)

**Proposition 145.** The exponential function \( \exp(z) \) has the following basic properties (not that we do not use the notation \( e^z \) here in order to reduce confusion with yet-undefined power functions):

(a) (Euler’s identity)

\[
\exp(i\pi) = -1.
\]

(b) \( \exp(z) \) is its own derivative.

(c) \( \exp(x + y) = \exp(x) \exp(y) \). Stated in another way, \( \exp \) is a homomorphism from the additive group of \( \mathbb{C} \) to the multiplicative group.

(d) The notation \( \exp(x) \) is consistent with iterated multiplication as defined in definition 525 (a iv), that is, \( \exp(n) = e \cdot \ldots \cdot e \) and for positive integers \( n \), \( \exp(n) = \) and \( \exp(-n) = \frac{1}{\exp(n)} \).

(e)

\[ \exp(z) = \frac{1}{\exp(-z)}. \]

(f) For real \( t \), \( e^t \) is a positive real number.

(g) \( \exp(z) = \exp(\overline{z}) \).

(h) For any \( c \in \mathbb{R} \), the function \( t \mapsto \exp(it) \) is a bijection between any half-open interval \([c, c + 2\pi)\) and the unit circle in \( \mathbb{C} \).

(i) \( t \mapsto \exp(t) \) is a bijection from \( \mathbb{R} \) to \([0, \infty)\).
(j) For any $c \in \mathbb{R}$, $\exp(z)$ is a bijection between the strip $S := \{a + bi : c \leq b < c + 2\pi\}$ and the complex plane $\mathbb{C} \setminus \{0\}$.

(k) $\exp(z)$ is $2\pi$-periodic.

(l) For nonnegative real $t \geq 0$ we have

$$\exp(t) = \lim_{n \to \infty} \left(1 + \frac{t}{n}\right)^n$$

Proof.

**Proof of 145 (a).** By eq. (63) and proposition 150, we have

$$\exp(i\pi) = \cos(\pi) + i\sin(\pi) = -1.$$

**Proof of 145 (b).** Follows from proposition 132 and corollary 124.

**Proof of 145 (c).** The Cauchy product of $\exp(x)$ and $\exp(y)$ is

$$\exp(x)\exp(y) = \left(\sum_{k=0}^{\infty} \frac{x^k}{k!}\right) \left(\sum_{k=0}^{\infty} \frac{y^k}{k!}\right) = \sum_{k=0}^{\infty} \sum_{m=0}^{k} \frac{x^m y^{k-m}}{m! (k-m)!} = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} x^m y^{k-m} \exp(x+y).$$

**Proof of 145 (d).** We use induction on $n$ to prove $\exp(n) = e^n$. The case $\exp(0) = 1$ is obvious. If we assume that $\exp(n) = e^n$, by proposition 145 (c), we have

$$\exp(n+1) = \exp(n)\exp(1) = e^n \cdot e = e^{n+1}.$$

Note that this works for negative $n$ too.

**Proof of 145 (e).** Note that

$$1 = \exp(0) = \exp(z - z) = \exp(z)\exp(-z),$$

hence

$$\exp(-z) = \frac{1}{\exp(z)}.$$

**Proof of 145 (f).** For $t > 0$, the following

$$\exp(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}.$$
is a series of positive real numbers. To see its convergence, we apply proposition 66 (d’Alambert’s ratio test):

\[
\frac{t^k}{k!} \cdot \frac{(k-1)!}{t^{k-1}} = \frac{t}{k} \to 0 \text{ as } k \to \infty.
\]

Thus, exp(t) is a nonnegative real number. Furthermore, since the sequence of partial sums is monotone, exp(t) cannot be zero. Hence, for \( t > 0 \), we have \( \exp(t) > 0 \).

**Proof of 145 (g).** By proposition 150,

\[
\exp(a + bi) = \exp(a) \exp(bi) = \exp(a) (\cos(b) + i \sin(b)) = \exp(a) (\cos(b) - i \sin(b)) = \exp(a) \exp(-bi) = \exp(a - bi) = \exp(a + bi).
\]

**Proof of 145 (k).** By proposition 145 (a),

\[
\exp(x + 2i\pi) = \exp(x) \exp(2i\pi) = \exp(x).
\]

Furthermore, this is also the minimal period. If we assume that sin(x) has another period, say \( p \in (0, 2\pi) \), we would have \( \sin(p) = \sin(0) = 0 \) and proposition 136 (a) would imply that \( \cos(p) \in \{-1, 1\} \). But then \( \cos(p) \) would be an extreme point for cos, which is not possible because cos is convex in \([0, 2\pi]\) and only has three extremal points — 0, \( \pi \), 2\pi.

**Proof of 145 (h).** For \( c, t \in \mathbb{R} \) we have

\[
|\exp(it)| = |\cos(t) + i \sin(t)| = \sqrt{\cos(t)^2 + \sin(t)^2} \stackrel{(49)}{=} 1.
\]

Furthermore, if \( r \) is another real number,

\[
\exp(ir) = \exp(i(t + (r - t))) = \exp(it) \exp(i(r - t)). \quad (66)
\]

It follows that \( \exp(ir) \neq \exp(it) \) if and only if \( \exp(i(r - t)) \neq 0 \). If \( t, r \in [c, c + 2\pi) \) and \( t \neq r \), this is satisfied.

Hence, \( t \mapsto \exp(it) \) is indeed an injection of \([c, c + 2\pi)\) into the unit circle of \( \mathbb{C} \). It is also a surjection because of the intermediate value theorem.

**Proof of 145 (i).** First, assume that \( e^t \) is not injective on \( \mathbb{R} \). Then there exist \( t, r \in \mathbb{R}, t \neq r \), such that \( e^t = e^r \). By proposition 145 (f), both are positive real numbers. In particular, we can divide by \( e^t \) to obtain

\[
1 = \frac{e^r}{e^t} \stackrel{(c)}{=} e^{r-t} \stackrel{(c)}{=} e^{r-t}.
\]
We know that $e^0 = 1$ from proposition 145 (d). Thus, it is enough to show that $e^t = 1$ if and only if $t = 0$.

Assume that $e^t = 1$ holds for some $t > 0$. The partial sums are monotonely increasing, so in order for them to converge to 1, for any fixed index $n$ we must have

$$0 \leq \sum_{k=0}^{n} \frac{t^k}{k!} = 1 + \sum_{k=1}^{n} \frac{t^k}{k!} \leq 1,$$

$$-1 \leq \sum_{k=1}^{n} \frac{t^k}{k!} \leq 0.$$

But $\sum_{k=1}^{n} \frac{t^k}{k!} > 0$ because $t > 0$. The obtained contradiction proves that $e^t \neq 1$ for positive $t$.

For negative $t$, note that

$$e^t e^{-t} = 1.$$

Since $-t$ is positive, $e^{-t} \neq 1$ and hence $e^t \neq 1$.

Therefore, the function $t \mapsto e^t$ is injective on $\mathbb{R}$. It is also surjective onto $\mathbb{R}^>0$ because of the intermediate value theorem.

**Proof of 145 (j).** Fix $a + bi \in S_c$, that is, $b \in [c, c + 2\pi)$. By proposition 145 (c),

$$e^{a+bi} = e^a e^{bi}.$$

By proposition 145 (h), $b \mapsto e^{bi}$ is injective for $b \in [c, c + 2\pi)$ and by proposition 145 (i), $a \mapsto e^a$ is injective on $\mathbb{R}$. It follows that their product is also injective.

**Proof of 145 (l).** By theorem 1287 (Newton’s binomial theorem),

$$\left(1 + \frac{t}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{t}{n}\right)^k 1^{n-k} =$$

$$= \sum_{k=0}^{n} \frac{n!}{(n-k)! k!} \frac{t^k}{n^k} =$$

$$= \sum_{k=0}^{n} \frac{n!}{(n-k)! k!} \frac{t^k}{n^k} =$$

$$= \sum_{k=0}^{n} \left[ \prod_{j=1}^{k} \left(1 - \frac{k + j}{n}\right) \right] \frac{t^k}{k!}.$$

Fix an index $m$. Since the series is nonnegative, there exists an index $N$ such that for $n \geq N$

$$\sum_{k=0}^{m} \frac{t^k}{k!} \leq \sum_{k=0}^{n} \left[ \prod_{j=1}^{k} \left(1 - \frac{k + j}{n}\right) \right] \frac{t^k}{k!}.$$

85
Note that
\[
\left[ \prod_{j=1}^{k} \left( 1 - \frac{k + j}{n} \right) \right] \frac{t^k}{k!} \leq \frac{t^k}{k!},
\]
hence
\[
\sum_{k=0}^{m} \frac{t^k}{k!} \leq \sum_{k=0}^{n} \left[ \prod_{j=1}^{k} \left( 1 - \frac{k + j}{n} \right) \right] \frac{t^k}{k!} \leq \sum_{k=0}^{n} \frac{t^k}{k!}.
\]
By lemma 55 (Squeeze lemma),
\[
\lim_{n \to \infty} \left( 1 + \frac{t}{n} \right)^n = \lim_{n \to \infty} \sum_{k=0}^{n} \frac{t^k}{k!} = \exp(t).
\]

**Definition 146.** Fix \( c \in \mathbb{R} \). Unless specified otherwise, we assume \( c = 0 \).

We define the **natural logarithm** \( \log(x) \) as inverse function of \( e^x \) from \( \mathbb{C} \setminus \{0\} \) to the strip \( S_c := \{ a + bi : c \leq b < c + 2\pi \} \).

We also define the **base b logarithm** \( \log_b(x) \) for \( b > 0 \) over the same domain as
\[
\log_b(x) := \frac{\log(x)}{\log(b)}.
\]

**Proof.** The well-definedness follows from proposition 145 (j). \( \square \)

**Proposition 147.**

(a) \( \log(xy) = \log(x) \log(y) \)

**Proof.**

Proof of 147 (a). Follows from proposition 145 (c). \( \square \)

**Definition 148.** For each positive real number \( y > 0 \), we define the **power function**
\[
x^y := e^{y \ln x}
\]
as a function of \( x \).

**Proposition 149.**

(a) \( (x^y)^z = x^{yz} \).

(b) \( D_x(x^y) = \log(x)x^y \).

**Proof.**

Proof of 149 (a).
\[
(x^y)^z = e^{z \log(e^{y \ln x})} = e^{zy \log(x)} = x^{yz}.
\]
Proof of 149 (b). Using the chain rule for differentiation, we obtain

\[ D_x(x^y) = D_x(e^{\log(y)x}) = \log(x)e^{\log(y)x} = \log(x)x^y. \]

Proposition 150. We have the following exponential-trigonometric identities:

(Euler’s formula) For any \( z \in \mathbb{C} \),

\[ e^{iz} = \cos(z) + i \sin(z). \] (67)

(Inverse Euler’s identities) For any \( z \in \mathbb{C} \),

\[
\begin{align*}
\sin(z) &= \text{real}(e^z) = \frac{e^{iz} - e^{-iz}}{2i} \quad (68) \\
\cos(z) &= \text{imag}(e^z) = \frac{e^{iz} + e^{-iz}}{2} \quad (69)
\end{align*}
\]

(De Moivre’s formula) For any complex number \( z \) and any nonnegative integer \( n \),

\[ (\cos(z) + i \sin(z))^n = \cos(nz) + i \sin(nz). \] (70)

Proof.

Proof of 150. Simply note that definition 144 is a termwise sum of definition 134 (a) and definition 134 (b), therefore eq. (67) holds.

Proof of 150. Follows from proposition 150.

Proof of 150. From proposition 150,

\[ (\cos(z) + i \sin(z))^n = e^{inz} \overset{149 (a)}{=} e^{i(nz)} = \cos(nz) + i \sin(nz). \]
3.6. Trigonometric polynomials

**Definition 151.** The ring of Laurent polynomials in the indeterminates \( \mathcal{X} \) over the integral domain \( D \) is obtained from the polynomial ring \( D[\mathcal{X}] \) by adjoining to \( D[\mathcal{X}] \) the set
\[
\left\{ \frac{1}{X} \mid X \in \mathcal{X} \right\}
\]
of reciprocals of the indeterminates from the field of rational functions \( D(\mathcal{X}) \).

If \( \mathcal{X} = \{X_1, \ldots, X_n\} \), this ring is denoted by \( R[X_1^\pm, \ldots, X_n^\pm] \) or \( R[X_1, X_1^{-1}, \ldots, X_n, X_n^{-1}] \). Individual polynomials are written as
\[
p(X_1, \ldots, X_n) = \sum_{k_1=-\infty}^{\infty} \cdots \sum_{k_n=-\infty}^{\infty} a_{k_1,\ldots,k_n} X_1^{k_1} \cdots X_n^{k_n}.
\]

**Definition 152.** We define the trigonometric polynomials over \( \mathbb{C} \) as the Laurent polynomials \( \mathbb{C}[e^{iz}] \). A trigonometric polynomial \( p \in \mathbb{C}[e^{iz}] \) can be written as
\[
p(z) = \sum_{k \in \mathbb{Z}} c_k e^{ikz}
\]
or, using Euler’s formula, rewritten in the more conventional notation (see [Boy08, p. 1] or [Rud87, p. 88]):
\[
p(z) = a_0 + \sum_{k=1}^{\infty} [a_k \cos(kz) + b_k \sin(kz)],
\]
where we denote \( a_k := c_k \) and \( b_k := ic_k \).

In particular, when using eq. (72), we may regard the coefficients \( \{a_k\}_{k=0}^\infty \) and \( \{b_k\}_{k=1}^\infty \) as either real or complex, which is a downside of eq. (71).

Denote by \( \tau_n(K) \) the vector space of all trigonometric polynomials of degree at most \( n \) with coefficients in \( K \). We also introduce the subspaces \( \tau_n^0 K \) of those polynomials which \( a_0 = 0 \).
3.7. Special functions

**Definition 153.** The Gamma function is

\[ \Gamma : \{ z \in \mathbb{C} \mid \text{real}(z) > 0 \} \to \mathbb{C} \]

\[ \Gamma(z) := \int_{0}^{\infty} x^{z-1} e^{-x} \, dx \]

**Theorem 154** (Stirling’s gamma approximation). For every nonnegative integer \( n \) there exists some constant \( \theta \in (0, 1) \) such that

\[ \Gamma(n + 1) = \sqrt{2\pi n} \cdot \left( \frac{n}{e} \right)^n \cdot e^{\frac{1}{12n + \theta}}. \]
3.8. Norms

Remark 155. Norms generalize distances of points in a plane, while absolute values generalize the absolute value over either $\mathbb{R}$ or $\mathbb{C}$. The axioms themselves differ minimally. Absolute values in a field are multiplicative norms over the field, however we cannot define absolute values in terms of norms since absolute values are needed for defining norms. Still, we will refer to fields with absolute values as normed fields.

Definition 156. Let $R$ be a semiring. We say that the function $|\cdot| : V \to \mathbb{R}_{\geq 0}$ is an absolute value or a semiring norm if

RN1 (identity) $x = 0_R$ if and only if $|x| = 0$

RN2 (multiplicativity) For any $x, y \in V$,

$$|xy| = |x| \cdot |y|$$

RN3 (subadditivity) For any $x, y \in V$,

$$|x + y| \leq |x| + |y|$$

Definition 157. Let $M$ be an $R$-module with absolute value $|\cdot|$. We say that the function $\|\cdot\| : M \to \mathbb{R}_{\geq 0}$ is a norm if

N1 (identity) $x = 0_M$ if and only if $\|x\| = 0_R$

N2 (absolute homogeneity)

$$\|tx\| = |t|\|x\| \text{ for all } t \in R \text{ and } x \in M$$

N3 (subadditivity)

$$\|x + y\| \leq \|x\| + \|y\| \text{ for all } x, y \in M$$

If we remove definition 157, then $\|\cdot\|$ is called a seminorm.
If instead $V$ is an associative and $\|\cdot\|$ satisfies the additional axiom

(a) (multiplicativity)

$$\|xy\| = \|x\| \cdot \|y\| \text{ for all } x, y \in M,$$

we say that it is a multiplicative norm.

Definition 158. A norm $\|\cdot\|$ on a real or complex vector space $V$ induces the metric

$$\rho : V \times V \to \mathbb{R}_{\geq 0}$$

$$\rho(x, y) := \|x - y\|.$$
Proof of M1. Follows from definition 157.

Proof of M2. By definition 157,
\[
\rho(x, y) = ||x - y|| = ||(-1)(y - x)|| = |-1||y - x|| = \rho(y, x).
\]

Proof of M3.
\[
\rho(x, y) + \rho(y, z) = ||x - y|| + ||y - z|| \geq ||x - z|| = \rho(x, z).
\]

Definition 159. We define the duality mapping
\[
D : E \rightrightarrows X^*,
D(x) := \{x^* \in X^* : ||x|| = ||x^*|| \text{ and } \langle x^*, x \rangle = ||x^*||||x||\}.
\]

We will usually use this mapping for unit vectors, so we may as well consider its restriction to the unit spheres, where
\[
D' : S_X \rightrightarrows S_{X^*},
D'(x) := \{x^* \in S_{X^*} : \langle x^*, x \rangle = 1\}.
\]

Definition 160. The norm ||·|| on X is called smooth if any of if for each \(x \in S_X\) the duality mapping is single-valued.

Definition 161. The norm ||·|| on X is called rotund or strictly convex if any of the following equivalent conditions hold:
(a) There are no line segments in the unit sphere \(S_X\).
(b) Every convex subset of X has at most one point of least norm.
(c) \(||x + y|| = ||x|| + ||y|| \implies x \text{ and } y \text{ are linearly dependent.} \quad (73)\)

Proof.

Proof that 161 (a) implies 161 (b). Let the norm in E be rotund and let \(C \subseteq E\) be a (potentially empty) convex set. We will prove that C contains at most one point of least norm.

If C is empty or otherwise contains no element of least norm, trivially contains at most one point of least norm.

Now let C contain at least one element \(x \in C\) of least norm. Assume that \(y \in C\) is another element of least norm. Necessarily \(||x|| = ||y||\).

Fix \(t \in (0, 1)\) and define \(z := tx + (1 - t)y\). Since C is convex, it contains z. Since x and y are elements of least norm, we have \(||z|| \geq ||x||\). By the triangle inequality,
\[
||z|| = ||tx + (1 - t)y|| \leq t||x|| + (1 - t)||y|| = ||x||,
\]
}\]
thus $\|z\| = \|x\|$.

This implies that the entire segment $[x, y]$ are elements of least norm in $C$. Hence, the segment $[x, y]$ is contained in the sphere $\|x\|S_E$, which contradicts the rotundity of the norm $\|\cdot\|$.

Hence, $C$ contains at most one element of least norm.

**Proof that 161 (b) implies 161 (a).** Let every convex set $C \subseteq E$ have at most one element of least norm.

Assume that the norm $\|\cdot\|$ is not rotund. Then the unit sphere $S_E$ contains a line segment $[x, y], x \neq y$. The set $[x, y]$ is compact and, by the Weierstrass extreme value theorem, the norm attains its minimum on the segment in a point $z \in [x, y]$. Since the segment is also convex and we assumed that convex sets have at most one element of least norm, it follows that this element $z$ is unique.

Then for any point $s \in [x, y], s \neq z$, we have $\|s\| > \|z\| = 1$, thus $s$ cannot be an element of the unit sphere. The obtained contradiction shows that the norm $\|\cdot\|$ is rotund.

**Proof that 161 (a) implies 161 (c).** Let $E$ be rotund let $x, y \in E$ be distinct vectors such that

$$\|x + y\| = \|x\| + \|y\|. \quad (74)$$

If either of them is the zero vector, then they are trivially linearly dependent. Assume that both $x$ and $y$ are nonzero and define

$$\xi := \frac{x}{\|x\|} \quad \eta := \frac{y}{\|y\|} \quad t := \frac{\|x\|}{\|x + y\|}$$

Equation (74) implies that

$$1 - t = 1 - \frac{\|x\|}{\|x + y\|} = \frac{\|x + y\| - \|x\|}{\|x + y\|} = \frac{\|y\|}{\|x + y\|}.$$ 

Since both $\xi$ and $\eta$ are in $S_E$, by rotundity, their convex combination

$$\nu := t\xi + (1 - t)\eta$$

should not be contained in $S_E$ unless $\xi = \eta$.

Calculating the norm, we obtain

$$\|\nu\| = \|t\xi + (1 - t)\eta\| = \|t\xi\| + \|(1 - t)\eta\| = \|\|x\|\xi\| \|\|y\|\eta\| = \|x + y\| = 1,$$

hence $\nu \in S_E$. Thus, $\xi = \eta$ and $x = \frac{\|x\|}{\|y\|}y$, so $x$ and $y$ are linearly dependent.
Proof that 161 (c) implies 161 (a). Let eq. (73) hold and fix \( x, y \in S_E, t \in (0, 1) \). Define \( z := tx + (1 - t)y \). First, assume that the vectors \( tx \) and \( (1 - t)y \) satisfy the left part of eq. (73), i.e.

\[
\|z\| = \|tx + (1 - t)y\| = t\|x\| + (1 - t)\|y\| = 1.
\]

This does not refute rotundity since \( x \) and \( y \) are not necessarily distinct. It follows from eq. (73) that \( tx \) and \( (1 - t)y \) are linearly dependent, hence \( x \) and \( y \) are also linearly dependent. Since \( x \) and \( y \) both have unit norm, either \( y = x \) or \( y = -x \).

If we assume that \( y = -x \), then

\[
\|z\| = \|tx + (1 - t)y\| = (2t - 1)\|x\| = 2t - 1,
\]

which is only possible if \( t = 1 \) since \( \|z\| = 1 \). But \( t \) is strictly less than 1.

Hence, \( y \neq -x \) and the only remaining possibility is that \( y = x \).

Now assume that the vectors \( tx \) and \( (1 - t)y \) do not satisfy the left part of eq. (73). This implies \( \|z\| < 1 \). Thus, \( x \) and \( y \) are necessarily distinct, but \( z \) is not contained in the unit sphere and the segment \([x, y]\) is not contained in \( S_E \).

We have shown that \( x, y \in S_E \) implies that either \( y = x \) or that the segment \([x, y]\) is not contained in \( S_E \), thus the norm in \( E \) is rotund. \( \square \)

Theorem 162. If the norm in a Banach space \( X \) is such that its dual norm in \( X^* \) is rotund (resp. smooth), then it is itself smooth (resp. rotund).

Proof. 1. First, let the dual norm \( \|\cdot\|^* \) be rotund and assume that \( \|\cdot\| \) is not smooth.

Fix \( x \in S_X \). Since \( D(x) \) is nonempty (by corollary 187) and since \( \|\cdot\| \) is not smooth, then there exist two different functionals \( x^*, y^* \in D(x) \), such that

\[
\langle x^*, x \rangle = \langle y^*, x \rangle = 1.
\]

We will show that the segment \([x^*, y^*]\) is contained in \( S_{X^*} \), i.e. that the dual norm is not rotund.

Fix any \( t \in (0, 1) \) and define \( z^* := tx^* + (1 - t)y^* \). We only need to show that \( \|z^*\| = 1 \).

By the triangle inequality, we have

\[
\|z^*\| = \|tx^* + (1 - t)y^*\| \leq t\|x^*\| + (1 - t)\|y^*\| = t + (1 - t) = 1.
\]

For the reverse inequality, note that

\[
\|z^*\| \geq \langle z^*, x \rangle = t\langle x^*, x \rangle + (1 - t)\langle y^*, x \rangle = t + (1 - t) = 1,
\]

thus \( \|z^*\| = 1 \). Hence, \([x^*, y^*]\) is contained in \( S_{X^*} \) and the dual space is not smooth. The obtained contradiction proves that the norm in \( X \) is rotund.
2. Now let the dual norm $\|\cdot\|^*$ be smooth and assume that $\|\cdot\|$ is not rotund. Then there exist points $x, y \in S_X$ such that the whole segment $[x, y]$ is contained in $S_X$.

Fix $t \in (0, 1)$ and define $z := tx + (1 - t)y \in S_X$. Denote by $J : X \to X^{**}$ the canonical embedding into the double-dual. By corollary 187, there exists a functional $z^* \in X^*$, such that

$$\langle J(z), z^* \rangle = \langle z^*, z \rangle = 1.$$ 

Because the dual norm $\|\cdot\|^*$ is smooth, we cannot have $\langle J(x), z^* \rangle = \langle z^*, x \rangle = 1$ or $\langle J(y), z^* \rangle = \langle z^*, y \rangle = 1$ and since $\|z^*\| = 1$, necessarily

$$\langle z^*, x \rangle < 1 \text{ and } \langle z^*, y \rangle < 1.$$ 

If follows that

$$1 = \langle z^*, z \rangle = t \langle z^*, x \rangle + (1 - t) \langle z^*, y \rangle < t + (1 - t) = 1,$$

which is a contradiction. Hence, $\|\cdot\|$ is rotund.

\[\square\]

**Proposition 163.** Norms in Hilbert spaces are both smooth and rotund.

**Proof.** Let $X$ be a Hilbert space, i.e. the norm is generated by an inner product and, due to Riesz’s theorem, we identify the space $X$ with its continuous dual $X^*$.

To prove that $X$ is rotund, choose $x, y \in S_X, x \neq y$. We will show that the segment $[x, y]$ is not contained in $S_X$.

If $x$ and $y$ are linearly dependent, necessarily $y = -x$ and all non-trivial convex combinations of $x$ and $y$ are contained in the open unit ball, hence $[x, y] \not\subseteq S_X$.

Not let $x$ and $y$ be linearly independent. By the Cauchy-Bunyakovskiy-Schwarz inequality, we have

$$\langle x, y \rangle \leq |\langle x, y \rangle| < \|x\|\|y\| = 1. \quad (75)$$

Fix $t \in (0, 1)$ and let $z := tx + (1 - t)y$. We will show that $z \not\in S_X$. Indeed,

$$\|z\|^2 = \langle z, z \rangle = t^2 \|x\|^2 + t(1 - t) \langle x, y \rangle + (1 - t)t \langle y, x \rangle + (1 - t)^2 \|y\|^2 =
\begin{align*}
t^2 + (1 - t)^2 + 2t(1 - t) \langle x, y \rangle < \\
(75) \quad t^2 + (1 - t)^2 + 2(1 - t) = \\
= t^2 + 1 - 2t + t^2 + 2t - t^2 = 1.
\end{align*}$$

Thus, $\|z\|^2 < 1$ and $\|z\| < 1$ and $z \not\in S_X$.

In both cases, no interior point of the segment $[x, y]$ is contained in $S_X$, hence the norm in $X$ is rotund.

Since we identify $X$ with its dual, the norm in $X^*$ is also rotund and by theorem 162, the norm in $X$ is also smooth. \[\square\]
Example 164. The norms in $c_0$ and $l^1$ are neither smooth nor rotund.

Proof. Consider the space $c_0$ of all real sequences that converge to zero equipped with the uniform norm

$$\|x\|_{c_0} := \sup_i |x_i|.$$  

Note that the dual space of $c_0$ is (isometrically isomorphic to) the space $l^1$ of absolutely summable sequences with norm

$$\|x\|_{l^1} := \sum_i |x_i|.$$  

Let $\{e_n\}_{n=1}^\infty$ be the canonical basis of $c_0$, i.e. the coordinates $e_n^{(i)}$ of $e_n$ are given by the Dirac delta function, $e_n^{(i)} := \delta_{i,n}$.

For every natural $n \geq 1$, define $x_n$ to be the same as $e_n$ except that the first coordinate of $x_n$ is always 1.

The corresponding norms of $e_n$ are all equal to 1 and the norms of $x_n$ are

$$\|x_n\|_{c_0} = 1 \quad \text{and} \quad \|x_n\|_{l^1} = 2.$$  

For every $n$ we have

$$\langle e_1, x_n \rangle = \langle e_n, x_n \rangle = 1,$$

hence $J_{c_0}(x_n)$ has at least two elements $e_1$ and $e_n$ and the norm in $c_0$ is not smooth.

Given that $\{x_1, x_2, \ldots\} \subseteq S_{c_0}$, consider the convex combinations of $x_2$ and $x_3$:

$$tx_2 + (1-t)x_3 = (1, t, (1-t), 0, 0, 0, \ldots).$$

Evidently $tx_2 + (1-t)x_3 \in S_{c_0}$ for every $t \in (0,1)$, hence the norm in $c_0$ is not rotund.

The contrapositions to the statements in theorem 162 say that if $X$ is not rotund (resp. smooth), then the dual space $X^*$ is not smooth (resp. rotund). Thus, $l^1$ is neither smooth or rotund as the dual of $c_0$.  \[\square\]
4. Functional analysis

Section 2 (Real analysis) and section 3 (Complex analysis) study certain functions with values in finite-dimensional Hilbert spaces. Functional analysis studies spaces of functions arising from real or complex analysis. These spaces are mostly infinite-dimensional, and a lot of results hold for general infinite-dimensional vector spaces.

Nevertheless, we will only study vector spaces over $\mathbb{R}$ or $\mathbb{C}$. This is justified by remark 401. As in section 3 (Complex analysis), through this section, $\mathbb{K}$ will refer to either $\mathbb{R}$ or $\mathbb{C}$. 
4.1. Topological groups

**Definition 165.** Let $G$ be any group and let $\mathcal{T}$ be a topology on $G$. The tuple $(G, \cdot, \mathcal{T})$ is called a **topological group** if the group structure and topological structure agree, that is, the operations $\cdot : X \times X \rightarrow X$ and $(-)^{-1} : X \rightarrow X$ are continuous with respect to $\mathcal{T}$.

See remark 166 and definition 167 for more nuances.

**Remark 166.** It is conventional to require the topology in a topological group to be $T_1$ (see definition 306). We will not do this due to our goal of not assuming more than is necessary.

Due to proposition 171, it is immaterial whether we require the topology to be $T_0$ or $T_{3.5}$ or anywhere in between. It is customary to call the space “Hausdorff” (although stronger separation axioms actually hold) and require $T_1$ to hold (since it is simple to state).

We will explicitly mention when we want a topological group to be Hausdorff. This is usually, so when we speak of convergence.

**Definition 167.** The category $\text{TopGrp}$ of topological groups is a subcategory of both $\text{Top}$ and $\text{Grp}$. Its morphisms are the **continuous** group homomorphisms.

**Proposition 168.** Fix $x, y \in G$ in a topological group $G$. If $U$ is a neighborhood of $x$, then both $V = yx^{-1}U$ and $W = Ux^{-1}y$ are neighborhoods of $y$.

**Proof.** Since the group operations are continuous, for fixed $x$ and $y$, the function $f(z) := xy^{-1}z$ is continuous.

Note that $U = f(V)$, hence $V$ is the preimage of $U$ under $f$ and it follows from the continuity of $f$ that $V$ is open.

Since $x \in U$, $yx^{-1}x = ye = y \in V$. Therefore, $V$ is a neighborhood of $y$.

The proof that $W$ is a neighborhood of $y$ is analogous. \qed

**Corollary 169.** In a topological group $G$, every neighborhood is a translation of a neighborhood of the origin $e$.

**Remark 170.** Corollary 169 provides a lot of uniformity by allowing us to only consider neighborhoods of zero when working with topological groups.

**Proposition 171.** If a topological group is $T_0$, it is automatically $T_{3.5}$.

**Proposition 172.** A Hausdorff topological group $G$ can be made into a uniform space by the families of entourages

$$V^I_A := \{(x, y) \in G \times G : x^{-1}y \in A\},$$

$$V^R_A := \{(x, y) \in G \times G : xy^{-1} \in A\},$$

where $A$ is a symmetric neighborhood of the origin $e$.

If $G$ is abelian, the two families of entourages coincide.

**Proposition 173.** If $\{a_\alpha\}_{\alpha \in \mathcal{X}}$ and $\{b_\alpha\}_{\alpha \in \mathcal{X}}$ are nets in a Hausdorff topological group $X$ that converge to $a$ and $b$, correspondingly, then $a_\alpha b_\alpha \rightarrow ab$.

**Proof.** Special case of proposition 179. \qed
4.2. Topological vector spaces

**Definition 174.** Let $X$ be any vector space and let $\mathcal{T}$ be a topology on $X$. The space $(X, +, \cdot, \mathcal{T})$ is called a **topological vector space** if the linear and topological structure agree, that is, the operations $+: X \times X \to X$ and $\cdot: X \times \mathbb{R} \to X$ are continuous with respect to $\mathcal{T}$.

Both the additive group $(X, +)$ and the multiplicative group $(X \setminus \{0\}, \cdot)$ are topological groups. We regard $X$ as a subgroup of its additive topological group.

See remark 166, definition 175 and definition 176 for more nuances.

Given that a topological vector space $X$ has both a topological and an algebraic structure, we should adapt certain definitions.

**Definition 175.** We define the **continuous dual space** $X^*$ of a topological space $X$ as the vector space of all continuous linear functionals. This differs drastically from definition 767 because in the general case, the continuous dual space may be trivial, i.e. only contain the zero functional. See remark 182.

We use the same notation for both the algebraic dual spaces and the continuous dual space because the meaning is usually clear from the context. In particular, hyperplanes as defined in definition 411 are only relevant to continuous linear functionals.

**Definition 176.** The category $\textbf{Top Vect}_K$ of topological vector spaces over $K$ is a subcategory of both $\textbf{Top}$ and $\textbf{Vect}_K$. Its morphisms are the continuous linear maps.

**Remark 177.** As in remark 170, we are only interested in neighborhoods of the origin $0$ since any neighborhood $U$ of $x$ is simply a translation of the neighborhood $U - x$ of the origin.

**Proposition 178.** A Hausdorff topological vector space $X$ is a uniform space with the families of entourages

$$V_A := \{(x, y) \in X \times X : x - y \in A\},$$

where $A$ is a symmetric neighborhood of the origin $0$.

**Proof.** Follows from proposition 172.

**Proposition 179.** If $\{a_\alpha\}_{\alpha \in \mathcal{X}}$ and $\{b_\alpha\}_{\alpha \in \mathcal{X}}$ are nets in a Hausdorff topological vector space $X$ that converge to $a$ and $b$, correspondingly, then

(a) $a_\alpha + b_\alpha \to a + b$.

(b) $\lambda a_\alpha \to \lambda a$ for any scalar $\lambda \in \mathbb{K}$.

**Proof.** Fix a neighborhood $U$ of $0$ and fix an index $\alpha_0$ such that for $\alpha \geq \alpha_0$ we have both $a - a_\alpha \in U$ and $b - b_\alpha \in U$.

**Proof of 179 (a).** For addition, we have

$$(a + b) - (a_\alpha + b_\alpha) = (a - a_\alpha) + (b - b_\alpha) \in 2U.$$
**Proof of 179 (b).** For scalar multiplication, we have

\[ \lambda a - \lambda a_x \in \lambda U. \]

In both cases the containing neighborhood does not depend on \( x \), hence the nets converge to their desired values. \( \square \)

**Corollary 180.** If \( f, g : X \to Y \) are continuous functions between topological vector spaces, then for any point \( x_0 \in X \) we have

\[
\lim_{x \to x_0} (f(x) + g(x)) = \lim_{x \to x_0} f(x) + \lim_{x \to x_0} g(x)
\]

and for any \( \lambda \in \mathbb{K} \)

\[
\lim_{x \to x_0} \lambda f(x) = \lambda \lim_{x \to x_0} f(x).
\]

**Definition 181.** We say that a topological vector space is **locally convex** if there exists a topological base of convex sets.

**Remark 182.** Given a Hausdorff locally convex space \( X \), corollary 185 shows that the canonical duality pairing as defined in remark 182 is nondegenerate. If the space is not locally convex, we cannot guarantee that the pairing will be nondegenerate and our restriction to continuous linear functionals could interfere with our habits of working with linear functionals.

**Definition 183.** We say that \( f : X \to \mathbb{R} \) is a **sublinear functional** if it satisfies

(a) (subadditivity) \( f(x + y) \leq f(x) + f(y) \) for any \( x, y \in X \).

(b) (positive homogeneity) \( f(tx) \leq tf(x) \) for any \( t > 0 \) and \( x \in X \).

Compare this definition to definition 543 (d).
4.3. The Hahn-Banach theorem

The Hahn-Banach theorem is an important result that can be stated differently and in different levels of generality.

**Theorem 184** (Geometric Hahn-Banach theorem/Mazur's theorem). Fix a topological vector space $X$. Let $A \subseteq X$ be an open convex set and $L \subseteq X$ be a subspace that is disjoint from $A$. Then there exists a continuous linear functional $x^* \in X^*$ such that

$$\text{real}(x^*, x) > 0, \quad x \in A$$
$$\text{real}(x^*, x) = 0, \quad x \in L$$

*See remark 43 for a justification of only considering the real part of $x^*$.*

**Corollary 185.** The dual of a Hausdorff locally convex space $X$ does not vanish at the nonzero vectors of $X$.

*Proof.* Fix a nonzero point $x \in X$. The result follows from theorem 184 (Geometric Hahn-Banach theorem/Mazur's theorem) with $L := \{0\}$ and $A$ — any convex set containing $x$ and not containing zero. Such a set $A$ exists because the topology is Hausdorff and $x$ has a neighborhood disjoint from any point in $L$. \hfill $\square$

**Corollary 186.** The annihilator of any proper subspace of a Hausdorff locally convex space contains nonzero elements.

*Proof.* Denote the proper subspace by $L \subsetneq X$. Fix $x \in X \setminus L$ and let $A$ be a convex neighborhood of $x$ that is disjoint from $L$. The result follows from theorem 184 (Geometric Hahn-Banach theorem/Mazur's theorem). \hfill $\square$

**Corollary 187.** In a normed space $X$, for any nonzero vector $x \in X$ there exists a continuous functional $x^* \in S_X^*$ such that $(x^*, x) = \|x\|$. In other words, the duality mapping is nonempty for any point.

*Proof.* This follows from corollary 186 by taking $A := B(x, \|x\|)$ and $L := \{0\}$ and then scaling the obtained functional. \hfill $\square$

**Theorem 188** (Hahn-Banach hyperplane separation theorem). Fix a topological vector space $X$. Let $A, B \subseteq X$ be disjoint convex sets. If $f A \neq \emptyset$, there exists a continuous linear functional separating $A$ and $B$. 


4.4. Frechet spaces

Definition 189. An F-space is a complete metrizable topological vector space. We can assume that an F-space is a tuple $(X, \rho)$, where $\rho$ is a complete translation-invariant metric. A Frechet space is a locally convex F-space.
4.5. Banach spaces

Definition 190. A Banach space is a normed vector space which is also a complete metric spaces with the metric induced by the norm.

Definition 191. Let $M$ and $N$ be left $R$-modules. A duality pairing $\langle \cdot , \cdot \rangle : M \times N \to R$ is a nondegenerate bilinear form.

The canonical duality pairing of a vector space $V$ over $F$ is

$$\langle \cdot , \cdot \rangle : V^* \times V \to F$$

$$\langle x^* , x \rangle \mapsto x^*(x).$$

[Mer17] Example 192. Consider the polynomial algebra $\mathbb{R}[x]$ as a vector space with the supremum norm. We will show that it is not complete. Define the sequence

$$p_n(x) := \sum_{k=0}^{n} \frac{x^k}{2^k}, n = 1, 2, \ldots$$

Then the limit of the sequence in $C([0, 1])$ is the power series

$$\lim_{n \to \infty} p_n(x) = \sum_{k=0}^{n} \frac{x^k}{2^k} = \frac{2}{2 - x}.$$

Since $\mathbb{R}[x]$ is a subspace of $C([0, 1])$, we conclude that $\mathbb{R}[x]$ has fundamental sequence, but we just demonstrated that its limit is not in $\mathbb{R}[x]$.

Definition 193. Fix two nonempty Banach spaces $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$. We define the operator norm $\| \cdot \|_{\text{hom}(X,Y)}$ on $\text{hom}(X,Y)$ equivalently as

(a) $$\|L\|_{\text{hom}(X,Y)} := \sup_{\|x\|_X = 1} \|Lx\|_Y.$$ 

(b) $$\|L\|_{\text{hom}(X,Y)} := \sup_{\|x\|_X < 1} \|Lx\|_Y.$$ 

(c) $$\|L\|_{\text{hom}(X,Y)} := \sup_{x \neq 0} \frac{\|Lx\|_Y}{\|x\|_X}.$$ 

(d) $$\|L\|_{\text{hom}(X,Y)} := \inf \{c \geq 0 : \|Lx\|_Y \leq c\|x\|_X\}.$$ 

In particular, this induces a norm on $X^*$. 

102
**Definition 194.** Let $X$ be a Banach space.

We define the **support function** $\sigma_{A^*}$ for the set of functionals $A^* \subseteq X^*$ by

$$
\sigma_{A^*} : X \to \mathbb{R} \cup \{\infty\} \\
\sigma_{A^*}(x) := \sup\{(x, x) : x^* \in A^*\}
$$

and the **weak* support function** $\sigma^*_A$ for the set of points $A \subseteq X$ by

$$
\sigma^*_A : X^* \to \mathbb{R} \cup \{\infty\} \\
\sigma^*_A(x^*) := \sup\{(x^*, x) : x \in A\}.
$$

**Definition 195.** Given a linear functional $x^*$, a nonempty subset $A$ of $X$ and a **diameter** $\alpha > 0$, the value $S(x^*, A, \alpha)$ is called a slice of $A$, where

$$
S : X^* \times \text{pow}(X) \times \mathbb{R}_{>0} \mapsto \text{pow}(A) \\
S(x^*, A, \alpha) := \{x \in A : \langle x^*, x \rangle > \sigma^*_A(x^*) - \alpha\}.
$$

We define a weak* slice of $A^* \subseteq X^*$ as $S^*(x, A^*, \alpha)$, where

$$
S^* : X^* \times \text{pow}(X) \times \mathbb{R}_{>0} \mapsto \text{pow}(A) \\
S^*(x, A^*, \alpha) := \{x^* \in A^* : \langle x^*, x \rangle > \sigma_{A^*}(x) - \alpha\}.
$$

If we need to make the underlying space explicit, we will use $S_X(x^*, A, \alpha)$ and $S^*_X(x, A^*, \alpha)$.

**Proposition 196.** If $\{a_k\}_{k=1}^\infty$ and $\{b_k\}_{k=1}^\infty$ are sequences in a **Banach algebra** $X$, that converge to $a$ and $b$, correspondingly, then $a_k b_k \to ab$.

**Proof.** Let $\delta > 0$ and let $k_0$ be an index such that for $k \geq k_0$ we have both $\|a - a_k\| < \delta$ and $\|b - b_k\| < \delta$. Then

\[
ab - a_k b_k = (ab - ab_k) + (ab_k - a_k b) + (a_k b - a_k b_k) = \\
= a(b - b_k) + (ab_k - ab + ab - a_k b) + (-a_k)(b_k - b) = \\
= a(b - b_k) + a(b_k - b) + (a - a_k)b + (-a_k)(b_k - b) = \\
= a \underbrace{(b - b_k)}_{\in B(0, \delta)} + a \underbrace{(b_k - b)}_{\in B(0, \delta)} + (a - a_k)b + (-a_k) \underbrace{(b_k - b)}_{\in B(0, \delta)}.
\]

Therefore, $\|ab - a_k b_k\| < \delta^2 + \|a + b\| \delta$. If we require $\delta$ to be strictly less than 1, we obtain $\delta^2 < \delta$ and $\|ab - a_k b_k\| < (1 + \|a + b\|)\delta$.

Given an arbitrary $\varepsilon > 0$, we can choose $\delta = \frac{\min\{\varepsilon, 1\}}{1 + \|a + b\|}$ in order to have $\|ab - a_k b_k\| < \varepsilon$ for some large enough $k$.

Therefore, $a_k b_k \to ab$. 

\[\Box\]
4.6. Hilbert spaces

Definition 197. A Hilbert space is an inner product space which is also a complete metric space with the metric induced by the inner product.

Definition 198. A set of vectors $A$ in a Hilbert space $X$ is called an orthonormal system if

$$\langle x, y \rangle := \begin{cases} 1, & x = y, \\ 0, & x \neq y. \end{cases}$$

It is a special case of an orthogonal system. We are usually interested in orthogonal bases.
4.7. Asplund spaces

**Definition 199.** The Banach space $X$ is called an Asplund (resp. weak Asplund) space if any of the following equivalent conditions hold:

1. Every continuous convex function on a convex open subset $D$ of $X$ is Frechet (resp. Gateaux) differentiable at a dense $G_δ$ subset of $D$.

2. The dual space $X^*$ has the Radon-Nikodym property.

3. Every nonempty weak* compact convex subset of $X^*$ is the weak* closed convex hull of its weak* strongly exposed points.
4.8. Minkowski functionals

**Definition 200.** Let $A$ is an absorbing convex set.

We define the corresponding **Minkowski functional**

$$
\rho_A : X \to [0, \infty), \quad \rho_A(x) = \inf\{t > 0 : x \in tA\}.
$$

**Proof.** We will prove that $\rho_A(x)$ is always a nonnegative real number. Obviously

$$
\rho_A(x) \geq 0
$$

since the infimum over $\mathbb{R}_{>0}$ is 0.

Now fix $x \in X$. Since $A$ is an absorbing set, there exists $t_0 > 0$ such that $t_0x \in A$. We need to take the infimum of all such numbers. This infimum exists since $\mathbb{R}$ is complete and the set over which we take the minimum is bounded. \qed
4.9. Dentable sets

**Definition 201.** A subset \( A \) of a Banach space \( X \) is called **dentable** if it admits slices of arbitrarily small diameter, i.e. for every \( \varepsilon > 0 \) there exist a functional \( x^* \in X^* \) and a diameter \( \alpha > 0 \), such that \( \operatorname{diam} S(x^*, A, \alpha) < \varepsilon \).

Weak* dentability is defined in an obvious way.

**Definition 202.** The space \( X \) is said to have the **Radon-Nikodym property (RNP)** if every nonempty bounded set \( A \) of \( X \) is dentable.

**Proposition 203.** Let \( X \) be a Banach space and \( A^* \subseteq X^* \) be a weak*-dentable set. Then \( A^* \) is dentable in \( X^* \).

**Proof.** Let \( \varepsilon > 0 \) and let \( x \in X \) and \( \alpha > 0 \) be such that \( \operatorname{diam} S(x, A^*, \alpha) < \varepsilon \). We denote by \( J(x) \) the embedding of \( x \in X \) into the double-dual \( X^{**} \) and by \( T(J(x), A^*, \alpha) \) the slice of \( A^* \) in \( X^* \). We have that

\[
S(x, A^*, \alpha) = \{ x^* \in A^* : \langle x^*, x \rangle > \sigma_{A^*}(x) - \alpha \} = \\
= \{ x^* \in A^* : \langle x^*, x \rangle > \sup\{ \langle y^*, x \rangle : y^* \in A^* \} - \alpha \} = \\
= \{ x^* \in A^* : \langle J(x), x^* \rangle > \sup\{ \langle J(x), y^* \rangle : y^* \in A^* \} - \alpha \} = T(J(x), A^*, \alpha),
\]

Since \( J \) is an isometry, this equality implies that

\[
\operatorname{diam} T(J(x), A^*, \alpha) = \operatorname{diam} S(x, A^*, \alpha) < \varepsilon.
\]

Hence, \( A^* \) admits arbitrarily small slices in \( X^* \), i.e. it is dentable in \( X^* \). \( \square \)
4.10. Differentiability

Let $X$ and $Y$ be Hausdorff topological vector spaces and let $U \subseteq X$ be an open set.

**Definition 204.** Our goal is to study the following partial operator:

$$
\partial : \text{Set}(U, Y) \times U \times X \to Y
$$

$$
\partial(f, x, h) := \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t}.
$$

We implicitly assume that $t \neq 0$ because otherwise the definition would not make sense. We only use the operator $\partial$ inside this definition. See remark 205 for a discussion of derivative notation.

The quotient under the limit sign is called a difference quotient.

For each function $f : U \to Y$, each point $x_0 \in X$ and each “direction” vector $x_0 \in X$, we want to obtain a value in $Y$, which we will call the directional derivative of $f$ at $x_0$ in the direction $h$. Note that $h$ is allowed to range over $X$.

The existence of a directional derivative is already a harsh condition, however we impose even harsher restrictions.

(a) If, for fixed $f$ and $x_0$, the directional derivative $\partial(f, x_0, h)$ exists for all directions $h$, we define the **first variation** of $f$ at $x_0$ as

$$
\delta f(x_0) : X \to Y
$$

$$
[\delta f(x_0)](h) := \partial(f, x_0, h).
$$

Within its domain of definition of $\delta$, which is stricter than that of $\partial$, the operator $\delta$ is a currying of $\partial$. We are interested in how the operator $\delta f(x)$ varies as $x$ varies.

Note that $\delta f(x_0)$ is an operator from $X$ to $Y$ even if $f$ is a function from $U \subseteq X$ to $Y$.

Note that, in general, the first variation operator $\delta f(x_0)$ is not linear - for example, by proposition 94, the first variation of a general convex function is, at most, sublinear. See section 2.9 (Nonsmooth derivatives) for how “nonlinear derivatives” are handled.

(b) If the first variation $\delta f(x_0)$ is a continuous linear operator, we say that $f$ is **Gateaux differentiable** or **weakly differentiable** at $x_0$ with **Gateaux derivative** $f'_G(x_0) := \delta f(x_0)$.

Since $f'(x_0)$ is linear in $h$, we can replace $t \downarrow 0$ with $t \to 0$ in eq. (76) and reformulate this condition of Gateaux differentiability as the existence of a continuous linear operator $\Lambda : X \to Y$ such that

$$
\Lambda(h) = \lim_{t \to 0} \frac{f(x_0 + h) - f(x_0)}{t}.
$$

If $\Lambda$ exists, we usually denote it by $D_G f(x_0)$ or $f'_G(x_0)$ and call it the **Gateaux derivative** of $f$ at $x_0$. See remark 205 for a discussion of the notation.
Wenowrestrictourattentionto Banach spaces. Wesay that \( f \) is **Frechet differentiable** or **strongly differentiable** at \( x_0 \) if there exists a continuous linear operator \( \Lambda : X \to Y \) such that
\[
\lim_{h \to 0} \frac{\|f(x_0 + h) - f(x_0) - \Lambda(h)\|_Y}{\|h\|_X} = 0.
\]
(79)

If \( \Lambda \) exists, we usually denote it by \( D_f(x_0) \) or \( f(x_0) \) and call it the **Frechet derivative** of \( f \) at \( x_0 \). See remark 205 for a discussion of the notation.

Note that eq. (78) uses convergence in the topology of \( Y \) while eq. (79) uses convergence in \( \mathbb{R} \). We discuss in remark 206 how Frechet differentiability is a special “uniform” case of Gateaux differentiability.

If there exists a continuous linear operator \( \Lambda \) such that
\[
\lim_{y \to x_0 \atop z \to x_0} \frac{\|f(y) - f(z) - \Lambda(y - z)\|_Y}{\|y - z\|_X} = 0,
\]
(80)
we say that \( f \) is **strictly differentiable** at \( x_0 \).

Remark 205. The following are standard notations for derivatives (some of the comments are based on [Fic68b, p. 146]):

(a) We already used **Lagrange’s notation** \( f'(x_0) \) and \( f_0'(x_0) \) in definition 204. Brevity is the only benefit of this notation. It becomes convenient when the functions have no name is a burden for directional derivatives.

The second and third derivatives of \( f \) at \( x_0 \) are denoted as \( f''(x_0) \) and \( f'''(x_0) \) and the \( n \)-th derivative of is denoted as \( f^{(n)} \).

See definition 88 for variations of this notation.

(b) Newton’s notation is similar to that of Leibniz, but depends on placing dots on top of \( f \), e.g. \( \dot{f}(x_0) := f''(x_0) \). This is used in areas like mathematical physics, however it has not become standard in more pure areas of analysis.

(c) We use **Euler’s notation** \( Df(x_0) := f'(x_0) \) for more complicated expressions, e.g. corollary 124. The main benefit of this notation is that is allows to express differentiation as an operator, similar to what we defined in definition 204. The directional derivative of \( f \) at \( x_0 \) in the direction \( h \) is denoted as \( D_h f(x_0) \). Iterated differentiation corresponds to the standard notation for group composition: the \( n \)-th derivative at \( x_0 \) is denoted as \( D^n f(x_0) \).

We also use other letters in the superscripts like \( D^G f(x_0) \) for Gateaux derivatives, \( D^p f(x_0) \) for Clarke’s generalized derivatives, etc.

(d) Some authors like [Phe93] use a variation of Euler’s notation with \( \delta \) instead of \( D \). For example, directional derivatives are introduced as \( \delta^+ f(x_0)(h) \) in [Phe93, lemma 1.2]. This is consistent with the standard notation for subdifferentials - see definition 103, however Euler’s notation appears to be more widely adopted.
(e) The Leibniz notation for the derivative \( f'(x_0) \) is
\[
\frac{df}{dx}(x_0) := Df(x_0).
\]
This notation is used extensively in integral calculus, however it is often confusing when manipulating derivatives. The fraction notation is unjustified in anything, but trivial cases and the partial derivative notation
\[
\frac{\partial f}{\partial x}(x_0) := D_x f(x_0)
\]
is even more confusing.
Note also that this depends on the convention of having variable names.

Remark 206. We will compare Gateaux differentiability with Frechet differentiability. Let \( X \) and \( Y \) be Banach spaces, let \( U \subseteq X \) be an open set and let \( f : U \rightarrow Y \) be an arbitrary function. Fix a point \( x_0 \in U \).

The continuous linear operator \( \Lambda : X \rightarrow Y \) is a Gateaux derivative if, for every \( \varepsilon > 0 \) and every direction \( h \in X \) there exists \( \delta_h^G > 0 \) such that
\[
\|\frac{f(x_0 + th) - f(x_0)}{t} - \langle \Lambda, h \rangle\|_Y < \varepsilon \quad \forall t \in (0, \delta_h^G).
\] (81)

In order for \( \Lambda \) to be a Frechet derivative, for every \( \varepsilon > 0 \) there must exist a \( \delta_F > 0 \), so that
\[
\|\frac{f(x_0 + h) - f(x_0) - \langle \Lambda, h \rangle}{\|h\|_X}\|_Y < \varepsilon \quad \forall h \in B(0, \delta_F) \setminus \{0\},
\]
which can be restated as
\[
\|\frac{f(x_0 + th) - f(x_0) - \langle \Lambda, h \rangle}{t}\|_Y < \varepsilon \quad \forall t \in (0, \delta_F) \forall h \in S_X.
\] (82)

By comparing eq. (81) to definition 204 (c), we conclude that \( f \) is Frechet differentiable at \( x_0 \) if \( \inf_{h \in S_X} \delta_h^G > 0 \), that is, if \( f \) is Gateaux differentiable and the convergence of the Gateaux derivative is uniform on \( h \in S_X \).

In particular, Frechet differentiability implies Gateaux differentiability.
4.11. Banach space interpolation

Definition 207. Let $\mathbb{K}$ be either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers.

(a) We say that two topological vector spaces $X_0$ and $X_1$ are **compatible** if they can both be embedded continuously into a Hausdorff topological vector space $\mathcal{U}$, in which case we can regard them as subspaces of $\mathcal{U}$.

In particular, both $X_0$ and $X_1$ are Hausdorff. We write $\overline{X} := (X_0, X_1)$.

(b) Denote by

$$\Delta \overline{X} := X_0 \cap X_1$$

the **intersection** of $X_0$ and $X_1$ (when regarded as subspaces of $\mathcal{U}$).

(c) Denote by

$$\Sigma \overline{X} := (X_0 + X_1)$$

the **sum** of $X_0$ and $X_1$. If $x \in \Sigma \overline{X}$, then there exist (possibly nonunique) vectors $x_0 \in X_0$ and $x_1 \in X_1$ such that $x = x_0 + x_1$.

(d) Let $\overline{X}$ be a pair of compatible spaces. We say that the space $X$ is an **intermediate** space for $\overline{X}$ if $\Delta \overline{X} \subseteq X \subseteq \Sigma \overline{X}$ with continuous linear inclusions.

(e) We introduce **morphisms** between two compatible pairs $\overline{X}$ and $\overline{Y}$ that are, strictly speaking, not functions between the pairs themselves. We define an **operator** $T : \overline{X} \rightarrow \overline{Y}$ between compatible pairs to be a function $T$ from $\Sigma \overline{X}$ to $\Sigma \overline{Y}$ that satisfies the additional conditions

$$T(X_0) \subseteq Y_0 \quad \text{and} \quad T(X_1) \subseteq Y_1.$$  

(f) If $\mathcal{C}$ is a subcategory of the category $\text{TopVect}_\mathbb{K}$ of topological vector spaces. We define the category $\text{Interp}_\mathbb{K}$ as the product category $\text{TopVect}_\mathbb{K} \times \text{TopVect}_\mathbb{K}$. More explicitly:

- **1116 (a)** The class of objects is the class of all pairs of compatible spaces.

- **1116 (b)** The morphisms between two compatible pairs are the continuous linear operators $T : \overline{X} \rightarrow \overline{Y}$ between them.

- **1116 (c)** Composition of morphisms is the usual function composition if we regard a morphism $T : \overline{X} \rightarrow \overline{Y}$ as a function from $\Sigma \overline{X}$ to $\Sigma \overline{Y}$.

- **1116 (g)** We say that the intermediate spaces $X$ for $\overline{X}$ and $Y$ for $\overline{Y}$ are a pair of **interpolation spaces** with respect to $\overline{X}$ and $\overline{Y}$ if, for any continuous linear operator $T : \overline{X} \rightarrow \overline{Y}$ between the compatible pairs, we have $T(X) \subseteq Y$.

Lemma 208. In a preordered magma $M$,

$$\max\{ab, cd\} \leq \max\{a, c\} \cdot \max\{b, d\}. \quad (83)$$
Proof. Since \( a \leq \max\{a, c\} \), then
\[
ab \leq \max\{a, c\} \cdot b \leq \max\{a, c\} \cdot \{b, d\}
\]
Analogously, \( cd \leq \max\{a, c\} \cdot \{b, d\} \) and
\[
\max\{ab, cd\} \leq \max\{a, c\} \cdot \{b, d\}.
\]
\[\square\]

**Proposition 209.** Let \( X := (X_0, X_1) \) be a compatible pair of Banach spaces.

(a) The intersection \( \Delta \mathcal{X} = X_0 \cap X_1 \) is a Banach space with norm
\[
\|x\|_{\Delta \mathcal{X}} := \max\{\|x\|_{X_0}, \|x\|_{X_1}\}.
\]

(b) The sum \( \Sigma \mathcal{X} = X_0 + X_1 \) is a Banach space with norm
\[
\|x\|_{\Delta \mathcal{X}} := \inf\{\|x_0\|_{X_0} + \|x_1\|_{X_1} : x_0 + x_1 = x\}.
\]

Proof. **Proof of 209 (a).** We will first show that \( \|x\|_{\Delta \mathcal{X}} \) is indeed a norm.

**N1** We have
\[
\|x\|_{\Delta \mathcal{X}} = \max\{\|x\|_{X_0}, \|x\|_{X_1}\} = 0 \text{ if and only if } \|x\|_{X_0} = \|x\|_{X_1} = 0.
\]
Clearly \( 0 \) belongs to both \( X_0 \) and \( X_1 \) hence to their intersection. Therefore, (86) is satisfied if and only if \( x = 0 \).

**N2** Absolute homogeneity follows from
\[
\|tx\|_{\Delta \mathcal{X}} = \max\{\|tx\|_{X_0}, \|tx\|_{X_1}\} = |t| \max\{\|x\|_{X_0}, \|x\|_{X_1}\} = |t|\|x\|_{\Delta \mathcal{X}}.
\]

**N3** Subadditivity follows from
\[
\|x + y\|_{\Delta \mathcal{X}} = \max\{\|x + y\|_{X_0}, \|x + y\|_{X_1}\} \leq \max\{\|x\|_{X_0} + \|y\|_{X_0}, \|x\|_{X_1} + \|y\|_{X_1}\} \leq \max\{\|x\|_{X_0}, \|x\|_{X_1}\} + \max\{\|y\|_{X_0}, \|y\|_{X_1}\} = \|x\|_{\Delta \mathcal{X}} + \|y\|_{\Delta \mathcal{X}}.
\]

We will now show the completeness of \( \|\cdot\|_{\Delta \mathcal{X}} \) directly. Let \( \{x_k\}_{k=1}^{\infty} \subseteq \Delta \mathcal{X} \) be a fundamental sequence. Both \( X_0 \) and \( X_1 \) are complete, therefore \( \{x_k\}_{k=1}^{\infty} \) converges to the same value. Both are subspaces of \( \mathcal{U} \), therefore the limit of the sequence is the same in both. In particular, it belongs to the intersection \( \Delta \mathcal{X} \).

Denote the limit of \( \{x_k\}_{k=1}^{\infty} \) by \( x \). Let \( \epsilon > 0 \) and let \( k_0 \) be an index such that both \( \|x_k - \xi_0\|_{X_0} < \epsilon \) and \( \|x_k - \xi_1\|_{X_1} < \epsilon \) whenever \( k \geq k_0 \). Then, for any \( k \geq k_0 \),
\[
\|x_k - x\|_{\Delta \mathcal{X}} = \max\{\|x_k - x\|_{X_0}, \|x_k - x\|_{X_1}\} < \epsilon.
\]
Therefore, the sequence \( \{x_k\}_{k=1}^{\infty} \) converges to \( x_0 \) in \( \Delta \mathcal{X} \).

112
Proof of 209 (b). Again, we will first show that \( \|x\|_{\Sigma X} \) is indeed a norm.

N1 Analogously to 209 (b),
\[
\|x\|_{\Sigma X} = \inf \{\|x_0\|_{X_0} + \|x_1\|_{X_1} : x_0 + x_1 = x \} = 0
\]
if and only if
\[
\|x\|_{X_0} = \|x\|_{X_1} = 0.
\]

N2 Absolute homogeneity follows from
\[
\|tx\|_{\Sigma X} = \inf \{\|tx_0\|_{X_0} + \|tx_1\|_{X_1} : x_0 + x_1 = x \} \leq |t| \|x\|_{\Sigma X}.
\]

N3 Subadditivity follows from
\[
\|x + y\|_{\Sigma X} = \inf (\|x_0\|_{X_0} + \|x_1\|_{X_1}) + (\|y_0\|_{X_0} + \|y_1\|_{X_1}) \leq \|x\|_{\Sigma X} + \|y\|_{\Sigma X}.
\]

It remains to prove the completeness of \( \|\cdot\|_{\Sigma X} \). Let \( \{x_0^{(k)} + x_1^{(k)}\}_{k=1}^{\infty} \subseteq \Sigma X \) be a fundamental sequence. Fix \( \epsilon > 0 \). Then there exists an index \( m_0 \) such that \( k, m \geq k_0 \) implies
\[
\|x_0^{(k)} + x_1^{(k)} - x_0^{(m)} - x_1^{(m)}\|_{\Sigma X} < \epsilon.
\]

But
\[
\|x_0^{(k)} - x_0^{(m)}\|_{X_0} \leq \| (x_0^{(k)} + x_1^{(k)}) - (x_0^{(m)} + x_1^{(m)})\|_{\Sigma X} < \epsilon,
\]
hence the sequence \( \{x_0^{(k)}\}_{k=1}^{\infty} \) is fundamental. Since \( X_0 \) is complete, this sequence has a limit, which we will denote by \( \xi_0 \). We define \( \xi_1 \) analogously.

With the same \( \epsilon \), denote by \( k_0 \) an index such that both \( \|x_0^{(k)} - \xi_0\|_{X_0} < \frac{\epsilon}{2} \) and \( \|x_1^{(k)} - \xi_1\|_{X_1} < \frac{\epsilon}{2} \) whenever \( k \geq k_0 \).

Then
\[
\| (x_0^{(k)} + x_1^{(k)}) - (\xi_0 + \xi_1)\|_{\Sigma X} = \inf \{\|x_0\|_{X_0} + \|x_1\|_{X_1} : x_0 + x_1 = (x_0^{(k)} + x_1^{(k)}) - (\xi_0 + \xi_1)\} \leq \|x_0^{(k)} - \xi_0\|_{X_0} + \|x_1^{(k)} - \xi_1\|_{X_1} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]

Therefore, \( \xi_0 + \xi_1 \) is the limit of the sequence \( \{x_0^{(k)} + x_1^{(k)}\}_{k=1}^{\infty} \subseteq \Sigma X \) in \( \Sigma X \). \( \square \)

Example 210. The spaces \( L^p(\mathbb{R}) \) are interpolation spaces for the pair \((L^1(\mathbb{R}), L^\infty(\mathbb{R}))\). The pair is compatible because both are subspaces of the space \( S(\mathbb{R}) \) of all Lebesgue-measurable real function with metric
\[
\rho(f, g) := \int_{\mathbb{R}} \frac{|f(x) - g(x)|}{1 + ||f(x) - g(x)||} d\lambda.
\]

113
Definition 211. [BL76, p. 6] Let \( \mu : U \to [0, \infty] \) be a positive measure and \( p \) be a positive real number. The **Lebesgue space** \( L_p \) is defined as the set of bounded functions \( f : U \mapsto \mathbb{K} \) such that the norm

\[
\|f\|_{L_p} := \begin{cases} 
\left( \int_U |f(t)|^p \, dt \right)^{1/p}, & 0 < p < \infty \\
\text{ess sup}_{t \in U} |f(t)|, & p = \infty
\end{cases}
\]

Theorem 212 (The Riesz-Thorin interpolation theorem). Fix two measure spaces \((U, \mu)\) and \((V, \nu)\). Let \( T : S(U, \mu) \to S(V, \nu) \) be a continuous linear map between the corresponding spaces of measurable functions.

Suppose that for some real numbers \( p_0, p_1, q_0, q_1 \geq 1 \) we have

\[
T(L_{p_j}(U, \mu)) \subseteq T(L_{q_j}(V, \nu)), \quad j = 0, 1.
\]

Additionally, let \( \theta \in (0, 1) \) and

\[
\frac{1}{p} = \frac{1 - \theta}{p_0} + \frac{\theta}{p_1}, \quad \frac{1}{q} = \frac{1 - \theta}{q_0} + \frac{\theta}{q_1}.
\]

Then

\[
T(L_p(U, \mu)) \subseteq L_q(V, \nu)
\]

and

\[
\|T\|_{\text{hom}(L_p, L_q)} \leq \|T\|_{\text{hom}(L_{p_0}, L_{q_0})}^{1-\theta} \|T\|_{\text{hom}(L_{p_1}, L_{q_1})}^\theta.
\]

Definition 213. [BL76, p. 6] Let \( \mu : U \to [0, \infty] \) be a positive measure and \( p \) be a positive real number.

(a) Given a scalar-valued function \( f : U \mapsto \mathbb{K} \), we define its **distribution function** as

\[
m_f : [0, \infty] \to \mathbb{K}, \quad m_f(\sigma) := \mu\{x : x > \sigma\}.
\]

(b) We define the **decreasing rearrangement** of \( f \) as

\[
f^*(t) := \inf\{\sigma : m_f(\sigma) \leq t\}.
\]

(c) The \((p, q)\)-Lorenz space, for potentially infinite positive \( q > 0 \), is the set of functions \( f : U \mapsto \mathbb{K} \) for which the quasinorm

\[
\|f\|_{L_{p,q}} := \begin{cases} 
\left( \int_0^\infty \left( \frac{f^*(\tau)}{\tau^p} \right)^q \frac{1}{\tau^q} \, d\tau \right)^{1/q}, & 1 \leq q < \infty \\
\text{ess sup}\left( \frac{f^*(t)}{t^p} \right), & q = \infty
\end{cases}
\]
is finite.
In particular, when \( q = \infty \), we use the notation
\[
\|f\|_{L^{p^*}} := \left( p \int_0^\infty \sigma^{p} m_f(\sigma) \frac{d\sigma}{\sigma} \right)^{\frac{1}{p}}
\]

**Theorem 214** (The Marcinkiewicz interpolation theorem). Fix two measure spaces \((U, \mu)\) and \((V, \nu)\). Let \( T : S(U, \mu) \to S(V, \nu) \) be a continuous linear map between the corresponding spaces of measurable functions.
Suppose that for some real numbers \( p_0, p_1, q_0, q_1 \geq 1 \) we have
\[
T(L^{p_j}(U, \mu)) \subseteq T(L^{q_j^*}(V, \nu)), j = 0, 1.
\]

Additionally, let \( \theta \in (0, 1) \) and
\[
\frac{1}{p} = \frac{1 - \theta}{p_0} + \frac{\theta}{p_1} \quad \quad \quad \quad \quad \frac{1}{q} = \frac{1 - \theta}{q_0} + \frac{\theta}{q_1}.
\]
Then, if \( p \leq q \),
\[
T(L^{p}(U, \mu)) \subseteq L^{q}(V, \nu)
\]
and
\[
\|T\|_{\text{hom}(L^{p}, L^{q})} \leq C \theta \|T\|_{\text{hom}(L^{p_0}, L^{q_0})}^{1-\theta} \|T\|_{\text{hom}(L^{p_1}, L^{q_1})}^\theta
\]
for some constant \( C \).

**Definition 215.** Let \( \overline{X} := (X_0, X_1) \) and \( \overline{Y} := (Y_0, Y_1) \) be compatible pairs of Banach spaces. If \( X \) and \( Y \) are a pair of interpolation spaces and, additionally, the inequality
\[
\|T\|_{\text{hom}(X, Y)} \leq C \|T\|_{\text{hom}(X_0, Y_0)}^{1-\theta} \|T\|_{\text{hom}(X_1, Y_1)}^\theta
\]
holds for some constant \( C > 0 \) and \( \theta \in [0, 1] \), we say that the pair \((X, Y)\) are **interpolation spaces of exponent** \( \theta \).
If, additionally, \( C = 1 \), we say that \((X, Y)\) is an **exact pair** of interpolation spaces.

**Definition 216.** Let \( \overline{X} := (X_0, X_1) \) be a compatible pair of Banach spaces. Instead of the norm \( \|\cdot\|_{X_1} \) in \( X_1 \), we can consider **equivalent norms** of the type \( t \|\cdot\|_{X_1} \) for \( t > 0 \). Furthermore, we can also introduce equivalent norms in \( \Sigma \overline{X} \) via the **K-functional**
\[
K : (0, \infty) \times \Sigma \overline{X} \quad K(t, x) := \inf \{ t \|x_0\|_{X_0} + t \|x_1\|_{X_1} : x_0 + x_1 = x \}.
\]
See proposition 217 \((c)\) for a proof that \( x \mapsto K(t, x) \) for a fixed \( t \geq 0 \) is an equivalent norm in the sum \( \Sigma \overline{X} \).

**Proposition 217.** The **K-functional** has the following basic properties:

(a) For any fixed \( x \in \Sigma \overline{X} \), the function \( t \mapsto K(t, x) \) is positive, **monotone** and **concave**.
(b) For positive real numbers \( t, s > 0 \), we have the following inequality:

\[
K(t, x) \leq \max\{1, \frac{t}{s}\}K(s, x). \tag{89}
\]

(c) For any fixed \( t > 0 \), the function \( x \mapsto K(t, x) \) is an equivalent norm in the sum \( \Sigma X \).

Proof.

Proof of 217 (a). That \( t \mapsto K(t, x) \) is positive is a slight generalization of definition 157, which can be proved as in proposition 209 (b).

Monotonicity follows from the monotonicity of the infimum.

To see that \( t \mapsto K(t, x) \) is concave, fix \( x, \lambda \in [0, 1] \) and \( t, s > 0 \). We have

\[
K(\lambda t + (1 - \lambda)s, x) = \inf\{\|x_0\|_{\Sigma X}, |x_1| \mid x_0 + x_1 = x\} = \inf\{\lambda \|x_0\|_{\Sigma X} + t|\|x_1\|_{\Sigma X} + (1 - \lambda)(\|x_0\|_{\Sigma X} + s|\|x_1\|_{\Sigma X})\mid x_0 + x_1 = x\} \geq \lambda K(t, x) + (1 - \lambda)K(s, x).
\]

Proof of 217 (b). Fix positive real numbers \( t, s > 0 \).

- If \( t \leq s \), by monotonicity we have

\[
K(t, x) \leq K(s, x) \tag{90}
\]

- If \( t > s \), we use concavity with

\[
s = \frac{s}{t}t + \left(1 - \frac{s}{t}\right)0
\]

to obtain

\[
K(s, x) \geq \frac{s}{t}K(t, x) + \left(1 - \frac{s}{t}\right)K(0, x).
\]

By positivity of \( K \), we have \( K(t, x) = 0 \) if and only if \( t = 0 \), hence

\[
K(t, x) \leq \frac{t}{s}K(s, x). \tag{91}
\]

Combining (90) and (91), we obtain (89).

Proof of 217 (c). That \( x \mapsto K(t, x) \) for a fixed \( t > 0 \) is a slight generalization of the proof in proposition 209 (b).

That the norms \( \|x\|_{\Sigma X} \) and \( K(t, \cdot) \) are equivalent follows from (89) with \( s = 1 \) for the upper bound and \( t = 1, s = t \) for the lower bound. That is,

\[
\min\{1, t\}K(1, x) \leq K(t, x) \leq \max\{1, t\}K(1, x).
\]

\[
\|x\|_{\Sigma X} \leq K(t, x) \leq \|x\|_{\Sigma X}
\]

\[\square\]
Example 218. The $K$-functional for the pair $(L_1(\mathbb{R}), L_\infty(\mathbb{R}))$ from example 210 is

$$K(t, f) := \int_0^t f^*(\tau) d\tau.$$ 

Definition 219. For $\vartheta \in \mathbb{R}$, $q \in (0, \infty]$ and nonnegative functions $g : [0, \infty) \to [0, \infty]$ we define

$$\Phi_{\vartheta, q}(g) := \begin{cases} \left( \int_0^\infty \left( \frac{g(\tau)}{\tau^\vartheta} \right)^q \frac{d\tau}{\tau} \right)^{\frac{1}{q}}, & 0 < q < \infty \\ \operatorname{ess sup}_{t \geq 0} \left( \frac{g(t)}{t^\vartheta} \right), & q = \infty \end{cases} \tag{92}$$

and

$$\gamma_{\vartheta, q} := \Phi_{\vartheta, q}(\min(t, 1)). \tag{93}$$

Proposition 220. The function $\Phi_{\vartheta, q}$ has the following basic properties:

(a) For $s > 0$ and $h(t) := g\left(\frac{t}{s}\right)$ we have

$$\Phi_{\vartheta, q}(h) = \frac{1}{s^\vartheta} \Phi_{\vartheta, q}(g). \tag{94}$$

(b) For finite $q$ we have

$$\gamma_{\vartheta, q} = \left( \frac{1}{q \vartheta(1 - \vartheta)} \right)^{\frac{1}{q}}. \tag{95}$$

Proof.

Proof of 220 (a). The case $q = \infty$ is obvious. For $0 < q < \infty$, we have

$$\Phi_{\vartheta, q}(h) = \left( \int_0^\infty \left( \frac{h(\tau)}{\tau^\vartheta} \right)^q \frac{d\tau}{\tau} \right)^{\frac{1}{q}} =$$

$$= \left( \frac{1}{s^\vartheta q} \int_0^\infty \left( \frac{g\left(\frac{\tau}{s}\right)}{\left(\frac{\tau}{s}\right)^\vartheta} \right)^q \frac{d\left(\frac{\tau}{s}\right)}{\left(\frac{\tau}{s}\right)} \right)^{\frac{1}{q}} =$$

$$= \frac{1}{s^\vartheta} \Phi_{\vartheta, q}(g).$$

Proof of 220 (b). We can raise $\gamma_{\vartheta, q}$ to the $q$-th power for brevity of notation:

$$\gamma_{\vartheta, q}^q = \Phi_{\vartheta, q}(\min(t, 1))^q =$$
\[
\begin{align*}
\int_0^1 \left( \frac{\tau}{\tau^\theta} \right)^q \frac{d\tau}{\tau} + \int_1^\infty \left( \frac{1}{\tau^\theta} \right)^q \frac{d\tau}{\tau} &= \\
= \int_0^1 \tau^{(1-\theta)q-1} d\tau + \int_1^\infty \tau^{-\theta q-1} d\tau &= \\
= \frac{1 - 0}{(1-\theta)q} + \lim_{\tau \to \infty} \tau^{-\theta q - 1} = \\
= \frac{1}{(1-\theta)q} - \frac{1}{-\theta q} = \\
= \frac{1}{(1-\theta)(-\theta)q} = \\
= \frac{1}{(1-\theta)q}
\end{align*}
\]

\[\square\]

**Definition 221.** Let \( \overline{X} := (X_0, X_1) \) be a compatible pair of Banach spaces.

For \( \theta \in (0, \infty), q \in (0, \infty) \), we introduce the following norm:

\[
\|x\|_{\theta,q,K} := \Phi_{\theta,q}(K(t, x)). \tag{96}
\]

The subspace of \( \overline{X} \) for which this norm is finite is denoted by either

\( \Sigma K_{\theta,q}(\overline{X}) \) or \( X_{\theta,q,K} \).

**Theorem 222.** Let \( \theta \in (0, 1) \) and \( q \in (0, \infty] \). The space \( X_{\theta,q,K} \) defined in eq. (96) is an exact interpolation space of exponent \( \theta \). Furthermore,

\[
K(s, x) \leq (\gamma_{\theta,q})^{-1}s^\theta \|x\|_{\theta,q,K}. \tag{97}
\]

**Proof.** Note that \( K(s, x) \) is a norm on \( \overline{X} \) by proposition 217 (c). Therefore, \( \|\cdot\|_{\theta,q,K} \), the composition of \( K(s, x) \) with the Lorenz quasinorm \( \Phi_{\theta,q} \), is a norm.

We denote by \( X_{\theta,q,K} \) the space consisting of all vectors from \( \overline{X} \) where the norm (96) is finite.

From (89) it follows that

\[
\min(1, \frac{t}{s}^\theta)K(s, x) \leq K(t, x)
\]

and hence

\[
\Phi_{\theta,q} \left( \min(1, \frac{t}{s}^\theta)K(s, x) \right) \leq \Phi_{\theta,q}(K(s, x)),
\]

depends on \( t \) norm in \( X_{\theta,q,K} \).

By (94), we have

\[
\Phi_{\theta,q}(\min(1, \frac{t}{s}^\theta)) = \frac{1}{s^\theta} \Phi_{\theta,q}(\min(1, t), \gamma_{\theta,q})
\]

and (97) follows.
It remains to show that $X_{\vartheta,q,K}$ is an exact interpolation space of exponent $\vartheta$. Note that $K(1,x) = \|x\|_{\Sigma X}$ and thus (97) with $s = 1$ implies that

$$\gamma_{\vartheta,q} \|x\|_{\Sigma X} \leq \|x\|_{\vartheta,q,K},$$

which shows that $X_{\vartheta,q,K}$ can be embedded continuously into $\Sigma X$.

On the other hand, for $x \in \Delta X$ we have

$$K(t, x) \leq \|x\| \leq \|x\|_{\Delta X}$$

and

$$K(t, x) \leq \|x\| \leq t\|x\|_{\Delta X}$$

since $x = x + 0$ and $K(t, x) \leq \|x\| \leq t\|x\|_{\Delta X}$ since $x = 0 + x$.

Therefore,

$$K(t, x) \leq \min(1, t)\|x\|_{\Delta X},$$

which after applying $\Phi_{\vartheta,q}$ becomes

$$\|x\|_{\vartheta,q,K} \leq \gamma_{\vartheta,q} \|x\|_{\Delta X}.$$

Hence, we have the chain of continuous linear inclusions of Banach spaces

$$\Delta X \subseteq X \subseteq \Sigma X.$$

Finally, to show that $X$ is an interpolation space of exponent $\vartheta$, fix a linear operator $T : \tilde{X} \mapsto \tilde{C}Y$ between compatible pairs and let $Y$ be an intermediate space for $\tilde{C}Y$. Then

$$K(t, Tx)_{\tilde{C}Y} \leq \inf\{\|y_0\|_{Y_0} + t\|y_1\|_{Y_1} : y_0 + y_1 = Tx\} \leq \inf\{\|T\|_{\text{hom}(X_0,Y_0)}\|x_0\|_{X_0} + t\|T\|_{\text{hom}(X_1,Y_1)}\|x_1\|_{Y_1} : x_0 + x_1 = x\} = \|T\|_{\text{hom}(X_0,Y_0)} K\left(\|T\|_{\text{hom}(X_1,Y_1)}, t, x\right).$$

By applying $\Phi_{\vartheta,q}$ to both sides and using (94), we obtain

$$\|Tx\|_{\tilde{Y}_{\vartheta,q,K}} \leq \|T\|_{\text{hom}(X_0,Y_0)} 1 - \vartheta \|T\|_{\text{hom}(X_0,Y_0)} \vartheta \|x\|_{\tilde{Y}_{\vartheta,q,K}}.$$

Thus, $X$ satisfies definition 215 for being an exact interpolating space with exponent $\vartheta$. □

**Definition 223.** For positive numbers $\vartheta$ and $q$, we denote by $\lambda^{\vartheta,q}$ the set of all doubly-infinite real sequences $\{x_k\}_{k=-\infty}^{\infty}$ such that the norm

$$\|\{x_k\}_{k=-\infty}^{\infty}\|_{{\lambda}^{\vartheta,q}} := \left( \sum_{k=-\infty}^{\infty} \left( \frac{|x_k|}{2k^\vartheta} \right)^q \right)^{\frac{1}{q}}$$

is finite.
Theorem 224. The vector $x \in \Sigma \mathcal{X}$ belongs to $X_{\varphi,q,K}$ if and only if the sequence $\{x_k\}_{k=-\infty}^{\infty}$ defined as

$$x_k := K(2^k, x)$$

belongs to $\lambda^{\varphi,q}$.

Furthermore, for any integer $k$ the following inequalities hold:

$$\frac{1}{2^q} \ln 2 \|x_k\|_{\lambda^{\varphi,q}} \leq \|x\|_{\varphi,q,K} \leq 2 \cdot \ln 2 \|x_k\|_{\lambda^{\varphi,q}}.$$

Proof. We have

$$\|x\|^{q}_{\varphi,q,K} = \int_{0}^{\infty} \left( \frac{K(\tau, x)}{\tau^q} \right)^q \frac{d\tau}{\tau} = \sum_{k=-\infty}^{\infty} \int_{2^k}^{2^{k+1}} \left( \frac{K(\tau, x)}{\tau^q} \right)^q \frac{d\tau}{\tau}.$$

By proposition 217 (b), for each integer $k$,

$$K(2^k, x) \leq 2K(2^k, x).$$

By the monotonicity of $K$, for $t \in [2^k, 2^{k+1}]$ we have

$$K(2^k, x) \leq K(t, x) \leq 2K(2^k, x).$$

Denote $x_k := K(2^k, x)$. For $2^k \leq t \leq 2^{k+1}$ we have

$$x_k \leq K(t, x) \leq 2x_k.$$

Therefore,

$$\|x\|^{q}_{\varphi,q,K} \leq \sum_{k=-\infty}^{\infty} \int_{2^k}^{2^{k+1}} \left( \frac{x_k}{\tau^q} \right)^q \frac{d\tau}{\tau} \leq 2^q \sum_{k=-\infty}^{\infty} \left( \frac{x_k}{2^{kq}} \right)^q \cdot \ln \frac{2^{k+1}}{2^k} = \ln 2 \cdot 2^q \sum_{k=-\infty}^{\infty} \left( \frac{x_k}{2^{kq}} \right)^q = 2^q \ln 2 \|\{x_k\}_{k=-\infty}^{\infty}\|_{\lambda^{\varphi,q}}$$

and similarly for the lower bound.

\[ \square \]

Definition 225. Let $\mathcal{X} = (X_0, X_1)$ be a compatible pair of Banach spaces. Let $x \in \Sigma \mathcal{X}$. Put

$$E : (0, \infty) \times \Sigma \mathcal{X}$$

$$E(t, x) := \inf\{\|x - x_0\|_{X_1} : \|x_0\|_{X_0} \leq t\}.$$ 

Proposition 226. When $\mathcal{X}$ are quasi-Banach spaces, the $E$-functional has the following basic properties:
(a) For fixed \( x \in \Sigma \bar{X} \), the function \( t \mapsto E(t, x) \) is decreasing.

(b) For \( \varepsilon \in (0, 1) \), we have

\[
E(t, x + y) \leq E(\varepsilon t, x) + E((1 + \varepsilon)t, y).
\]

(c) \( x = 0 \) if and only if \( E(t, x) = 0 \) for all \( t > 0 \).

(d) \[
E(t, x) = \sup \left\{ \frac{K(s, x) - t}{s} : s > 0 \right\}.
\]

**Definition 227.** Let \( \bar{X} = (X_0, X_1) \) be a compatible pair of Banach spaces. We define an approximation space \( E_{\alpha,q}(\bar{X}) \) for \( x \in \Sigma \bar{X} \) as the space of all members of \( \Sigma \bar{X} \) for which the following norm

\[
\| x \|_{\alpha,q,E} := \Phi_{-\alpha,q}(E(t, a)) \quad (102)
\]

is finite.

Here \( \alpha \) and \( q \) are both positive real numbers and \( q \) is potentially \( \infty \).

**Theorem 228.** Let \( X \) be a compatible pair of Banach spaces. Let \( \alpha \) and \( q \) be positive real numbers and define

\[
\theta := \frac{1}{\alpha + 1}, \quad r := \theta q.
\]

Then

\[
(E_{\alpha,\theta q}(\bar{X}))^\theta = K_{\theta q}(\bar{X}).
\]

**Theorem 229.** Let \( X \) be a compatible pair of Banach spaces. Let \( \theta, \alpha_0, \alpha_1, r_0, r_1 \) and \( q \) be positive real numbers such that \( \alpha_0 \neq \alpha_1 \) and define \( r := \theta q \) and

\[
\alpha := (1 - \theta)\alpha_0 + \theta \alpha_1, \quad \beta := -\frac{\alpha_1 - \alpha}{\alpha_0 - \alpha}.
\]

Then

\[
K_{\theta,q}(E_{\alpha_0, r_0}(\bar{X}), E_{\alpha_1, r_1}(\bar{X})) = E_{\alpha,q}(\bar{X})
\]

and

\[
E_{\beta,r}(E_{\alpha_0, r_0}(\bar{X}), E_{\alpha_1, r_1}(\bar{X}))^\theta = E_{\alpha,q}(\bar{X}).
\]
5. **Approximation theory**

Approximation theory studies how real-valued functions can be approximated by more well-behaved functions. In its modern form, these include inequalities and optimization in function spaces.
5.1. Lagrange polynomials

Definition 230. Given distinct elements \(x_0, \ldots, x_n\) of the field \(\mathbb{K}\), we form the polynomial

\[
\omega(X) := \prod_{k=0}^{n} (X - x_k).
\]

Proposition 231. For the polynomial \(\omega\) from definition 230, for \(k = 0, \ldots, n\) we have

\[
\omega'(x_j) = \prod_{j=0}^{n} (x_j - x_k),
\]

where \(\omega'\) is the algebraic derivative of \(\omega\).

Proof. Fix \(k \in \{0, \ldots, n\}\) and denote

\[
q(X) := \prod_{j=0}^{n} (X - x_j).
\]

Then

\[
\omega(X) = (X - x_k)q(X)
\]

so

\[
\omega'(X) = [q(X) + Xq'(X)] - x_kq'(X) = q(X) + (X - x_k)q'(X).
\]

So for \(x_k\) we have

\[
\omega'(x_k) = q(x_k) = \prod_{j=0}^{n} (x_k - x_j).
\]

\[\square\]

Theorem 232 (Lagrange interpolation). Let \(x_0, x_1, \ldots, x_n\) be pairwise distinct elements of \(\mathbb{K}\) and let \(y_0, y_1, \ldots, y_n\) be arbitrary elements of \(\mathbb{K}\). Then there exists a unique polynomial \(L(X)\) of degree at most \(n\) such that \(L(x_k) = y_k\) for \(k = 1, \ldots, n\).

Proof.

Proof of uniqueness. Suppose that \(p, q\) are polynomials of degree at most \(n\) that both satisfy \(L(x_k) = y_k\) for \(k = 1, \ldots, n\). Their difference \(p - q\) is a polynomial of degree at most \(n\) that has \(n + 1\) roots. By proposition 663 (c), \(p - q = 0\).

Proof of existence. We will construct the polynomial explicitly. Define the Lagrange basis polynomial

\[
L(X) = \sum_{m=0}^{n} y_m \prod_{j=0}^{n} \frac{(X - x_j)}{(x_m - x_j)}.
\]
For \( k = 0, 1, \ldots, n \) we have

\[
L(x_k) = y_k \prod_{j=0}^{n} \frac{(x_k - x_j)}{(x_k - x_j)} + \sum_{m=0}^{n} y_m \prod_{j=0}^{0} \frac{(x_k - x_m)}{(x_k - x_m)} \prod_{j=0}^{n} \frac{(x_k - x_j)}{(x_m - x_j)} = y_k.
\]

Therefore, \( L \) is the desired polynomial.

\[\text{Theorem 233 (Finite field Lagrange interpolation).} \quad \text{For a multivariate function} \ f : \mathbb{F}_q^n \rightarrow \mathbb{F}_q \ \text{over the finite field} \mathbb{F}_q, \ \text{there exists a unique multivariate polynomial} \ L(X_1, \ldots, X_n) \ \text{such that}
\]

- For any sequence of values \( x_1, \ldots, x_n \),

\[f(x_1, \ldots, x_n) = L(x_1, \ldots, x_n).\]

- For every monomial \( X_1^{\gamma_1} \cdots X_n^{\gamma_n} \) of \( L \), \( \gamma_i < q \) for \( i = 1, \ldots, n \).

\[\text{Proof.} \quad \text{For any point} \ (x_1, \ldots, x_n) \in \mathbb{F}_q^n, \ \text{the characteristic polynomial}
\]

\[c(X_1, \ldots, X_n) := \prod_{i=0}^{n} \left( \prod_{m=0}^{q-1} \frac{X_i - m}{x_i - m} \right)\]

satisfies

\[c(y_1, \ldots, y_n) = \begin{cases} 1, & \text{if} \ x_i = y_i \ \text{for all} \ i = 1, \ldots, n \\ 0, & \text{otherwise}. \end{cases}\]

As in \text{theorem 232} (Lagrange interpolation), give us the desired polynomial, a linear combination of these basis polynomials with coefficients corresponding to the values of \( f \) give us the desired polynomial. \qed
5.2. Bernstein inequalities

**Definition 234.** Consider the operator $T : C([a, b]) \to C([a, b])$.

(a) If $f([a, b]) \subseteq [0, \infty)$ implies $T(f)([a, b]) \subseteq [0, \infty)$, we say that $T$ is **positive**.

(b) If $f(x) \leq g(x)$ for all $x \in [a, b]$ implies $T(f)(x) \leq T(g)(x)$ for all $x \in [a, b]$, we say that $T$ is **monotone**.

**Definition 235.** We denote by $\tilde{C}([a, b])$ the subspace of $C([a, b])$ consisting of all continuous functions in $[a, b]$ which are periodic with minimal period $b - a$.

**Definition 236.** We introduce two operators.

(a) The **algebraic approximation error**

$$E_n : C([a, b]) \to [0, \infty]$$

$$E_n(f) := \inf_{p \in \pi_n} \| f - p \|.$$

(b) The **trigonometric approximation error**

$$\tilde{E}_n : C([a, b]) \to [0, \infty]$$

$$\tilde{E}_n(f) := \inf_{p \in \tau_n} \| f - p \|.$$

**Theorem 237 (Jackson’s trigonometric theorem).** For $f \in \tilde{C}[-\pi, \pi]$ we have

$$E_n(f) \leq \frac{6^{k+1}}{n^k} \omega \left( f^{(k)}, \frac{1}{n} \right).$$

**Theorem 238 (Szegö’s inequality).** For any nonnegative integer $n$ and any $s \in S_{\tau_n}$, we have

$$|s'(\theta)|^2 + n^2s^2(\theta) \leq n^2 \quad \forall \theta \in [-\pi, \pi]. \quad (103)$$

**Proof.** Fix $n = 1, 2, \ldots$ and $\alpha \in [-1, 1]$.

For brevity, denote $c(\theta) := \cos(n\theta)$. Let $\theta_s$ and $\theta_c$ be numbers in $[-\pi, \pi]$ such that

$$s(\theta_s) + c(\theta_c) = \alpha.$$

We will show that

$$|s'(\theta_s)| \leq |c'(\theta_c)| \quad (104)$$

This will, in turn, imply that

$$[s'(\theta_s)]^2 \leq [c'(\theta_c)]^2 = n^2[\sin(n\theta_c)]^2 \leq n^2[1 - \cos(n\theta_c)]^2 = n^2[1 - s(\theta_s)]$$

which is equivalent to (103).
Now we will prove (104). If $\theta_c$ is a critical point of $c$, i.e. if $r'(\theta_c) = 0$, then $c'(\theta_c) = s'(\theta_c)$ and (104) holds. Suppose that $\theta_c$ is not a critical point. Denote by

$$\theta_n := \frac{k}{n} \pi, k = -n, -n+1, \ldots, n-2, n-1.$$ 

the extrema of $c(\theta)$ in $[-\pi, \pi]$.

Define the auxiliary function

$$r(\theta) := c(\theta) - s(\theta - \theta_c + \theta_s).$$

We now have $r'(\theta_c) = c'(\theta_c) - s'(\theta_s) = 0$.

Furthermore, since $|s| = 1$, then $|s(\theta)| \leq 1$ for all $\theta \in [-\pi, \pi]$. Therefore, $r(\theta_k) \leq 0$ for all odd $k$ and $r(\theta_k) \geq 0$ for all even $k$.

If $\theta_c$ coincides with any of the extrema $\theta_k$, then $r(\theta_c)$ holds trivially. Suppose that $\theta_c$ is between $\theta_{k-1}$ and $\theta_k$. Without loss of generality, assume that $k$ is even.

By the intermediate value theorem, there exists a zero of $r$ between $r(\theta_c)$ and $r(\theta_k)$. If $r(\theta_c) < 0$, then $\theta_c$ is a local minimum and hence there exists a point $\theta_c'$ between $r(\theta_{k-1})$ and $\theta_c$ such that $r(\theta_{c'}) = r(\theta_c) = 0$. But this would imply that $r$ has more than $2n$ different roots in the interval $[-\pi, \pi]$, which is a contradiction. 

**Corollary 239** (Bernstein's trigonometric inequality). For any nonnegative integer $n$ and any $s \in \tau_n$ we have

$$|s'(\theta)| \leq n|s| \ \forall \theta \in [-\pi, \pi].$$

*(105)*

**Proof.** The case $n = 0$ is trivial. If $n > 0$, for any $s \in \tau_n$, we can apply (103) to $\frac{s}{|s|}$ to obtain

$$|s'(\theta)|^2 + n^2 s^2(\theta) \leq ||s||^2 n^2 \ \forall \theta \in [-\pi, \pi].$$

*(106)*

*(105)* follows directly. 

**Theorem 240** (Bernstein's trigonometric theorem). Let $f \in \tilde{\mathcal{C}}[-\pi, \pi]$ and

$$E_n(f) \leq \frac{A}{n^{k+\alpha}} \ \ n = 0, 1, 2, \ldots,$$

where $A \in \mathbb{R}$ and $\alpha \in (0, 1)$.

Then $f \in C^{(k)}[-\pi, \pi]$ and $f^{(k)}$ is $\alpha$-Hölder.

**Proof.** Since $E_n(f)$ is bounded by $\frac{A}{n^{k+\alpha}}$ on a compact interval, there exists a sequence $\{s_k\}_{k=0}^{\infty}$ such that

$$||f - s_n|| \leq An^{k+\alpha}.$$

Define the sequence of polynomials

$$v_j := \begin{cases} s_1, & j = 0, \\ s_{2j} - s_{2j-1} & j > 0. \end{cases}$$

126
It is now clear that

$$\|f - \sum_{j=0}^{n} v_j\| \xrightarrow{j \to \infty} 0$$

because

$$|f(\theta) - \sum_{j=0}^{n} v_j(\theta)| = |f(\theta) - s_{2n}(\theta)| \leq Ar^{k+\alpha}.$$ 

For each term of the series, we have

$$|v_j(\theta)| \leq |f(\theta) - s_{2j}(\theta)| + |f(\theta) - s_{2j-1}(\theta)| \leq \frac{A}{2^{(j)(k+\alpha)}} + \frac{A}{2^{(j-1)(k+\alpha)}}.$$ 

By setting $B := A(2^{k+\alpha})$, we obtain

$$|v_j(\theta)| \leq \frac{B}{2^{j(k+\alpha)}}.$$ 

By corollary 239 (Bernstein’s trigonometric inequality),

$$\|v_j^{(r)}\| \leq 2^{jr}\|v_j\| \leq \frac{B}{2^{(j-r)(k+\alpha)}}.$$ 

Then

$$\sum_{j=0}^{\infty} v_j^{(r)}(\theta)$$

converges uniformly, therefore

$$f^{(r)}(\theta) = \sum_{j=0}^{\infty} v_j^{(r)}(\theta).$$

\[\square\]

**Theorem 241** (Bernstein’s algebraic inequality). For any nonnegative integer $n$ and any $p \in \pi_n$ and $x \in (a, b)$ we have

$$|p'(x)| \leq n \frac{1}{(b - a)(b - x)}\|p\|.$$ 

**Theorem 242** (Bernstein’s algebraic theorem). Let $f \in C[a, b]$ and

$$E_n(f) \leq \frac{A}{n^{k+\alpha}} \quad n = 0, 1, 2, \ldots,$$

where $A \in \mathbb{R}$ and $\alpha \in (0, 1)$. Then $f \in C^{(k)}(a, b)$ and $f^{(k)}$ is $\alpha$-Hölder in every $[a_1, b_1]$ such that $a_1 > a$ and $b_1 < b$. 

127
6. General topology

Analysis and geometry which are mostly concerned with real vector spaces. Topology is concerned with the bare necessities in which desirable properties of these spaces. General topology, also called point-set topology, allows studying topological spaces via their individual points. This is in stark contrast with section 14 (Category theory) and algebraic topology, which study topological spaces via continuous functions.
6.1. Topological spaces

**Definition 243.** Let $X$ be any set and $\mathcal{T} \subseteq \text{pow}(X)$ be a family of subsets of $X$. $\mathcal{T}$ is called a topology on $X$ and the tuple $(X, \mathcal{T})$ is said to be a topological space if the following axioms are satisfied:

- $O_1 \ \emptyset, X \in T$
- $O_2 \ U, V \in T \implies U \cap V \in T$
- $O_3 \ \mathcal{T}' \subseteq T \implies \bigcap \mathcal{T}' \in T$

If the topology is obvious from the context, we say that $X$ is a topological space.

Elements of the set $X$ are called points of the topological space, elements of $\mathcal{T}$ are called open sets and set-theoretic complements of open sets are called closed sets.

If $x \in U \in T$, we say that $U$ is a neighborhood of $x$. Note that some authors (e.g. [Kel55, p. 38]) alternatively define neighborhoods as arbitrary sets that contain an open set that contains $x$. For simplicity, we define the subfamily

$$\mathcal{T}(x) := \{U \in T : x \in U\}.$$ 

We say that $U$ is a punctured neighborhood of $x$ if $U \cup \{x\}$ is an open set and, consequently, a neighborhood of $x$.

Dually, we can define the family $\mathcal{F}$ of closed sets, where

- $F_1 \ \emptyset, X \in F$
- $F_2 \ U, V \in F \implies U \cup V \in F$
- $F_3 \ \mathcal{F}' \subseteq F \implies \bigcup \mathcal{F}' \in F$

If $(X, \mathcal{T})$ is a topological space, we denote the corresponding family of closed sets by

$$\mathcal{F}_T := \{X \setminus U : U \in T\}.$$ 

**Definition 244.** On a space $X$, we can explicitly define the following standard topologies:

- (a) The discrete topology $\mathcal{T} := \text{pow}(X)$.
- (b) The indiscrete topology $\mathcal{T} := \{\emptyset, X\}$.
- (c) For any cardinal $\xi$, the co-$\xi$ topology $\mathcal{T} := \{A \subseteq X : \text{card } A < \xi\}$ and, in particular, cofinite ($\xi = \aleph_0$) and cocountable ($\xi = c$) topologies.

For a deeper connection between discrete and indiscrete topologies, see example 1192 (a).

**Proposition 245.** A set $A$ is open if and only if every point of $A$ has a neighborhood $U$ such that $U \subseteq A$.

**Proof.** This holds vacuously for empty sets. Assume that $A \subseteq X$ is nonempty.
**Proof of sufficiency.** Assume that \( A \) is open and let \( x_0 \in A \). Then \( A \) is a neighborhood of \( x_0 \) and the theorem holds trivially.

**Proof of necessity.** Assume that every point \( x \in A \) has a neighborhood \( U_x \) such that \( U_x \subseteq A \). Take the union

\[
B := \bigcup_{x \in A} U_x.
\]

Obviously \( B \subseteq A \). Aiming at a contradiction, suppose that \( y_0 \in A \setminus B \). Then \( y_0 \) has a neighborhood \( U_{y_0} \) such that \( U_{y_0} \setminus B \) is nonempty. But this is impossible by the definition of \( B \). The obtained contradiction proves \( B = A \). \( \square \)

**Remark 246.** It is sometimes easier to define a topology \( \mathcal{T} \) via a subset of \( \mathcal{T} \). We will gradually construct a topology from a bare family of sets in \( X \). First, we will give two definitions for a base, one on which does not require an existing topology.

**Definition 247.** Fix a topological space \((X, \mathcal{T})\). We say that the family \( \mathcal{B} \subseteq \mathcal{T} \) is a base for the topology \( \mathcal{T} \) if \( \mathcal{B} \) satisfies any of the equivalent conditions:

(a) Every open set \( U \in \mathcal{T} \) is the union \( U = \bigcup \mathcal{B}' \) of some subset \( \mathcal{B}' = \mathcal{B} \)

(b) For any point \( x \in X \) and for any neighborhood \( U \) of \( x \) there exists a set \( V \in \mathcal{B} \) in the base such that \( x \in V \subseteq U \)

**Proof.**

**Proof that 247 (a) implies 247 (b).** Fix a point \( x \in X \) and a neighborhood \( U \in \mathcal{T} \) of \( x \). Let \( \mathcal{B}' \) be a subfamily of \( \mathcal{B} \) such that

\[
U = \bigcup \mathcal{B}'.
\]

Then \( x \in V \) for at least one \( V \in \mathcal{B}' \).

**Proof that 247 (b) implies 247 (a).** Fix an open set \( U \in \mathcal{T} \). Then for every \( x \in U \), there exists a set \( V_x \in \mathcal{B} \) such that \( x \in V_x \subseteq U \). We have

\[
\bigcup_{x \in U} V_x \subseteq U \subseteq \bigcup_{x \in U} V_x,
\]

thus

\[
U = \bigcup_{x \in U} V_x.
\]

**Proposition 248.** Let \( X \) be an arbitrary set and let \( \mathcal{B} \) be a family of subset that satisfies

\[
\begin{align*}
B1 & \quad \bigcup \mathcal{B} = X \\
B2 & \quad \forall U, V \in \mathcal{B}, \forall x \in U \cap V, \exists W \in \mathcal{B} : x \in W \subseteq U \cap V
\end{align*}
\]

[Eng89] page 12
Then the family
\[ \mathcal{T} := \left\{ \bigcup B' : B' \subseteq B \right\} \]  
(107)
is a topology on \( X \). Furthermore, \( \mathcal{B} \) is a base of \( \mathcal{T} \).

In particular, the base on any topology satisfies proposition 248 – proposition 248.

Proof. We will first prove that \( \mathcal{T} \) is indeed a topology.

O1 \( \emptyset = \bigcup \emptyset \in \tau \) and \( X = \bigcup B \in T \) (by proposition 248)

O3 Fix \( \mathcal{T}' = \{ U_\alpha : \alpha \in A \} \subseteq T \). By definition 247 (a), every set \( U_\alpha \) has a corresponding subfamily \( \mathcal{B}_\alpha \) of \( \mathcal{B} \) such that \( U_\alpha = \bigcup \mathcal{B}_\alpha \).
Define \( \mathcal{B}' := \bigcup_{\alpha \in A} \mathcal{B}_\alpha \). Obviously \( \mathcal{B}' \subseteq B \) and thus, by proposition 248, \( \bigcup \mathcal{B} \in T \).

O2 Fix \( U, V \in T \) and families \( \mathcal{B}_U, \mathcal{B}_V \subseteq B \) such that \( U = \bigcup \mathcal{B}_U \) and \( V = \bigcup \mathcal{B}_V \).
Fix arbitrary \( U' \in B_U \) and \( V' \in B_V \). We will show that \( U' \cap V' \in \tau \).

By proposition 248, for every \( x \in U' \cap V' \) there exists a neighborhood \( W_x \) of \( x \) such that \( W \subseteq U' \cap V' \).
The family \( \mathcal{B}_{U',V'} := \{ W_x : x \in U' \cap V' \} \) is a subfamily of \( \mathcal{B} \) and thus \( U' \cap V' = \bigcup \mathcal{B}_{U',V'} \in T \).
Hence, by definition 243, \( U \cap V \in \tau \).

Now, for any \( U \in T \), by eq. (107), there exists a subfamily \( \mathcal{B}' \subseteq B \) such that
\[ U = \bigcup \mathcal{B}' \cdot \]
Hence, \( \mathcal{B} \) is a base for \( \mathcal{T} \).

Definition 249. We define the weight of \((X, \mathcal{T})\) as the cardinal
\[ w((X, \mathcal{T})) := \min\{|\mathcal{B}| : \mathcal{B} \text{ is a base for } \mathcal{T}\}. \]
We simply write \( w(X) \) when the topology is clear from the context.
Spaces for which \( w(X) \leq \aleph_0 \) are said to be second-countable.

Proof. The definition is correct because cardinals are well-ordered when regarded as initial ordinals.

Definition 250. Fix a topological space \((X, \mathcal{T})\). We say that the family \( P \subseteq T \) is a subbase for the topology \( \mathcal{T} \) if the family
\[ \mathcal{B} := \left\{ \bigcap P' : P' \text{ is a nonempty finite subset of } P \right\} \]
of finite intersections of \( P \) is a base of \( \mathcal{T} \).

131
Proposition 251. Fix a set $X$ and a family of subsets $P \subseteq \text{pow}(X)$. The family $P' := P \cup X$ is then a subbase of some topology on $X$.

Definition 252. Fix a topological space $(X, \mathcal{T})$ and a point $x \in X$. We say that the family $\mathcal{B}(x) \subseteq T$ is a local base for $\mathcal{T}$ at $x$ if every neighborhood of $x$ contains a set from $\mathcal{B}(x)$.

Given a base $\mathcal{B}$, unless explicitly noted, we consider the subfamily $\mathcal{B}(x)$ of all members of $\mathcal{B}$ containing $x$.

The indexed family of local bases $\{\mathcal{B}(x) : x \in X\}$ is called a neighborhood system of $\mathcal{T}$.

Proposition 253. Analogously to definition 247(a), a set $A$ containing $x$ is a neighborhood of $x$ if and only if $A$ is a union of elements of the local base $\mathcal{B}(x)$.

Proof. Analogous to the proof of the equivalence in definition 247.

Proposition 254. Let $X$ be an arbitrary set and let $\{\mathcal{B}(x) \subseteq \text{pow}(X) : x \in X\}$ be an indexed family of families of subsets of $X$ that satisfies

BP1 For every $x \in X$, $\mathcal{B}(x) \neq \emptyset$ and $x \in U$ for every $U \in \mathcal{B}(x)$.

BP2 For every $x \in X$ and for all $U, V \in \mathcal{B}(x)$, $\exists W \in \mathcal{B}(x) : W \subseteq U \cap V$.

BP3 For all $x, y \in X$, $x \in U \in \mathcal{B}(y)$ implies that there exists $V \in \mathcal{B}(x)$ such that $U \subseteq V$.

Then the family

$$\mathcal{B} := \bigcup_{x \in X} \mathcal{B}(x)$$

is the base of some topology $\mathcal{T}$ on $X$. Furthermore, $\{\mathcal{B}(x) \subseteq \text{pow}(X) : x \in X\}$ is a neighborhood system for $(X, \mathcal{T})$.

In particular, the local base on any topology satisfies proposition 254 – proposition 254.

Definition 255. We define the character of the point $x \in X$ as the cardinal

$$\chi(x) := \min\{\text{card } \mathcal{B}(x) : \mathcal{B}(x) \text{ is a local base for } \mathcal{T} \text{ at } x\}.$$ 

We define the character of $(X, \mathcal{T})$ as

$$\chi((X, \mathcal{T})) := \sup\{\chi(x) : x \in X\}.$$ 

We simply write $\chi(X)$ when the topology is clear from the context.

Spaces for which $\chi(X) \leq \aleph_0$ are said to be first-countable.

Proof. The definition is correct because cardinals are well-ordered when regarded as initial ordinals.

Definition 256. Combining definition 250 and definition 252, we define a local subbase for $\mathcal{T}$ at $x$ to be a family $P(x) \subseteq T$ such that every neighborhood $U$ of $x$ contains a finite intersection of sets from $P(x)$.

Given a subbase $P$, unless explicitly noted, we consider the subfamily $P(x)$ of all members of $P$ containing $x$. 
Definition 257. Let \((X, \mathcal{F})\) be a topological space. Define the closure operator

\[
\text{cl} : \text{pow}(X) \to \text{pow}(X)
\]

\[
\text{cl}(A) := \bigcap \{F : F \in \mathcal{F}, A \subseteq F\}.
\]

Proposition 258. The closure operator has the following basic properties

(a) The set \(A\) is closed if and only if \(A = \text{cl}A\).

(b) For any \(x \in X\), \(x \in \text{cl}A\) if and only if every neighborhood of \(x\) intersects \(A\).

(c) \(\text{cl}\) is monotone, i.e. if \(A \subseteq B\), then \(\text{cl}(A) \subseteq \text{cl}(B)\).

Proof.

Proof of 258 (a). The condition \(A = \text{cl}A\) is equivalent to \(A\) being a closed superset of itself, which is equivalent to \(A\) being closed.

Proof of 258 (b). Note that this proof relies on definition 264, however we do not use this property when defining the boundary.

Proof of sufficiency. Fix \(x \in \text{cl}A\) and let \(U\) be a neighborhood of \(x\). If \(x \in A\), then obviously \(x \in U \cap A \neq \emptyset\). If \(x \notin A\), then \(U \cap A \neq \emptyset\) by definition 264 (b). In both cases, we obtain \(U \cap A \neq \emptyset\), which proves the statement.

Proof of necessity. Fix \(x \in X\) and assume that every neighborhood of \(x\) intersects \(A\). Since the case \(x \in A\) is trivial, suppose that \(x \notin A\). By proposition 245, every neighborhood \(U\) of \(x\) does not entirely belong to \(A\). By definition 264 (b), \(x \in \text{fr} \subseteq \text{cl}A\).

Proof of 258 (c). If \(A \subseteq B\), every closed superset of \(B\) is also a closed superset of \(A\).

Proposition 259. Let \(X\) be an arbitrary set and let \(\text{cl} : \text{pow}(X) \to \text{pow}(X)\) be a function that satisfies

\[
\begin{align*}
\text{CO1} \quad &\text{cl}(\emptyset) = \emptyset \\
\text{CO2} \quad &\forall A \in \text{pow}(X), A \subseteq \text{cl}(A) \\
\text{CO3} \quad &\forall A, B \in \text{pow}(X), \text{cl}(A \cup B) = \text{cl}(A) \cup \text{cl}(B) \\
\text{CO4} \quad &\forall A \in \text{pow}(X), \text{cl}(\text{cl}(A)) = \text{cl}(A)
\end{align*}
\]

Then the family

\[
\mathcal{J} := \{X \setminus F : F = \text{cl}(F)\}
\]

is a topology on \(X\). Furthermore, \(\text{cl} = \text{cl}^\mathcal{J}\), where \(\text{cl}^\mathcal{J}\) is the closure operator on \((X, \mathcal{J})\).

In particular, the closure operator on any topology satisfies proposition 259 – proposition 259.

Definition 260. Let \((X, \mathcal{F})\) be a topological space. Define the interior operator

\[
\text{int} : \text{pow}(X) \to \text{pow}(X)
\]

\[
\text{int}(A) := \bigcup \{U : U \in \mathcal{T}, U \subseteq A\}.
\]
Proposition 261. The interior operator has the following basic properties

(a) A set \( A \) is a topological space is open if and only if \( A = \text{int}(A) \).

(b) \( \text{int} \) is monotone, i.e. if \( A \subseteq B \), then \( \text{int}(A) \subseteq \text{int}(B) \).

Proof.

Proof of 261 (a). Follows from proposition 258 (a) and proposition 262.

Proof of 261 (b). Follows from proposition 258 (c) and proposition 262. \( \square \)

Proposition 262. For every set \( A \subseteq X \) we have

- \( X \setminus \text{int}(A) = \text{cl}(X \setminus A) \)
- \( X \setminus \text{cl}(A) = \text{int}(X \setminus A) \)

Proof. Any open subset \( U \subseteq A \) is a closed superset of \( X \setminus A \). A point belongs to \( \text{int}(A) \) if it belongs to at least one open subset of \( A \), which happens if and only if it belongs to at least one closed superset of \( X \setminus A \). Therefore,

\[
X \setminus \text{int}(A) = X \setminus \bigcup \{ U : U \in T, U \subseteq A \} = \\
= X \setminus \bigcup \{ F : F \in F_T, X \setminus A \subseteq F \} \overset{X \setminus A = A}{=} \\
= \bigcup \{ F : F \in F_T, F \subseteq A \} = \\
= \text{cl}(A).
\]

The other equality is obtained by noting that \( X \setminus \text{cl}(A) = X \setminus (X \setminus \text{int}(A)) = \text{int}(A) \). \( \square \)

Proposition 263. Let \( X \) be an arbitrary set and let \( f : \text{pow}(X) \rightarrow \text{pow}(X) \) be a function that satisfies

- \( iO1 \) \( \text{int}(X) = X \)
- \( iO2 \) \( \forall A \in \text{pow}(X), \text{int}(A) \subseteq A \)
- \( iO3 \) \( \forall A, B \in \text{pow}(X), \text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B) \)
- \( iO4 \) \( \forall A \in \text{pow}(X), \text{int}(\text{int}(A)) = \text{int}(A) \)

Then the family

\[
\mathcal{T} := \{ U : U = \text{int}(U) \}
\]

is a topology on \( X \). Furthermore, \( f = f_{\mathcal{T}} \), where \( f_{\mathcal{T}} \) is the interior operator on \( (X, \mathcal{T}) \).

In particular, the interior operator on any topology satisfies proposition 263 – proposition 263.

Definition 264. For a subset \( A \) of a topological space we define its boundary \( \text{fr}(A) \) equivalently as

(a) \( \text{fr}(A) := \text{cl}(A) \setminus \text{int}(A) \)
(b) \( \text{fr}(A) \) is the set of all points \( x \in X \) such that every neighborhood of \( x \) intersects both \( A \) and \( X \setminus A \).

**Proof.** The equivalence of the definitions is trivial when \( \text{fr}(A) = \emptyset \). We assume that \( \text{fr}(A) \neq \emptyset \).

**Proof that 264 (a) implies 264 (b).** Let \( x \in \text{cl}(A) \setminus \text{int}(A) \).

Aiming for a contradiction, suppose that there is a neighborhood \( U \) of \( x \) that does not intersect \( A \). Then \( U \subseteq X \setminus A \). Hence, \( A \subseteq X \setminus U \). Since \( X \setminus U \) is closed, it follows that \( \text{cl}(A) \subseteq X \setminus U \) as the intersection of all closed supersets of \( A \). But \( x \notin X \setminus U \), therefore \( x \notin \text{cl}(A) \), which contradicts our choice of \( x \in \text{cl}(A) \).

This proves that every neighborhood of \( x \) intersects \( A \).

By passing to complements, we can reuse this to prove that every neighborhood of \( x \) intersects \( X \setminus A \) using proposition 262.

**Proof that 264 (b) implies 264 (a).** Suppose that every neighborhood of \( x \in \text{fr}(A) \) intersects both \( A \) and \( X \setminus A \). Therefore, no neighborhood of \( x \) is contained in neither \( A \) nor \( X \setminus A \). Hence,

\[
x \in (X \setminus \text{int}(X \setminus A)) \setminus \text{int}(A) = \text{fr}(X \setminus A).
\]

\[\square\]

**Proposition 265.** The topological boundary has the following basic properties

(a) \( \text{fr}(A) \) is a closed set.

(b) If \( \text{fr}(A) \) is not empty, it is not an open set.

(c) \( \text{fr}(A) = \text{fr}(X \setminus A) \).

**Proof.**

**Proof of 265 (a).** Note that

\[
\text{fr}(A) = \text{cl}(A) \setminus \text{int}(A) = \text{cl}(A) \cap (X \setminus \text{int}(A)),
\]

which is the intersection of two closed sets. Hence, \( \text{fr}(A) \) is itself a closed set.

**Proof of 265 (b).** Note that \( \text{fr}(A) \) is either empty or is not open because definition 264 (b) is incompatible with proposition 245.

**Proof of 265 (c).** By proposition 262,

\[
\text{fr}(A) = \text{cl}(A) \setminus \text{int}(A) = \text{cl}(A) \cap (X \setminus \text{int}(A)) \overset{262}{=} (X \setminus \text{int}(X \setminus A)) \cap \text{cl}(X \setminus A) = \text{cl}(X \setminus A) \setminus \text{int}(X \setminus A) = \text{fr}(X \setminus A).
\]

\[\square\]
Definition 266. Let \((X, \mathcal{T})\) be a topological space.

(a) We say that the point \(x_0 \in X\) is a **cluster point** or an **accumulation point** of the set \(A \subseteq X\) if \(x \in \text{cl}(A \setminus \{x\})\). It is not necessary for \(x_0\) to belong to \(A\).

(b) The set of all cluster points of \(A\) is called the **derived set** of \(A\) and is denoted by \(\text{derived}(A)\).

(c) If a set equals its derived set, we call it a **perfect set**.

(d) Points in \(A \setminus \text{derived}(A)\) are said to be **isolated points** of \(A\).

(e) If \(\text{derived}(A) = \emptyset\), that is, if \(A\) consists of only discrete points, we say that \(A\) is a **discrete set**.

Proposition 267. Derived sets have the following basic properties

(a) \(x\) is a cluster point of \(A\) if and only if every neighborhood of \(x\) intersects \(A \setminus \{x\}\).

(b) \(x\) is an isolated point of \(A\) if and only if there exists a neighborhood of \(x\) that does not intersect \(A \setminus \{x\}\).

(c) \(\text{derived}(A)\) is a closed set.

(d) \(A \cup \text{derived}(A) = \text{cl}(A)\).

(e) A set is closed if and only if it contains all of its cluster points. Compare this result to proposition 279.

(f) A set if closed if and only if every point is either a cluster point or an isolated point.

Proof.

Proof of 267 (a). If every neighborhood \(U\) of \(x \in A\) intersects \(A \setminus \{x\}\), by proposition 258 (b), \(x \in \text{cl}(A \setminus \{x\})\) and \(x\) is therefore a cluster point.

Proof of sufficiency of ??.. Dual to proposition 267 (a).

Proof of 267 (c). Consider the complement of \(\text{derived}(A)\). If it is empty, \(\text{derived}(A)\) is trivially closed. Otherwise, let \(x \in X \setminus \text{derived}(A)\).

- If \(x\) is an isolated point of \(A\), by proposition 267 (b) there exists a neighborhood of \(x\) that does not intersect \(A \setminus \{x\}\).

- If \(x\) is not a point of \(A\), aiming at a contradiction, assume that every neighborhood of \(x\) intersects \(A\). Then, by definition 264 (b), \(x \in \text{fr}(A)\). But \(\text{fr}(A) \subseteq \text{cl}(A)\) and \(\text{cl}(A) = \text{cl}(A \setminus \{x\})\) because \(x\) does not belong to \(A\). Therefore, \(x\) is a cluster point of \(A\). This contradicts our assumption that \(x \notin \text{derived}(A)\), hence we can conclude that there exists a neighborhood of \(X\) that does not intersect \(A = A \setminus \{x\}\).
In both cases, proposition 245 allows us to conclude that $X \setminus \text{derived}(A)$ is open and, hence, derived$(A)$ is closed.

**Proof of 267 (d).** Clearly $A \subseteq \text{cl}(A)$. Also
\[
\text{derived}(A) \subseteq \bigcup_{x \in X} \text{cl}(A \setminus \{x\}) \subseteq \text{cl}(A).
\]

Now we will prove the reverse inclusion. Let $x \in \text{cl}(A)$. Then either $x \in A$ or $x \in \text{fr}(A)$. Assume the latter. By definition 264 (b), every neighborhood $U$ of $x$ has points both in $A$ and outside $A$, therefore $U \cap (A \setminus \{x\})$ is nonempty. By proposition 258 (b), $x \in \text{cl}(A \setminus \{x\})$, that is, $x \in \text{derived}(A)$.

**Proof of 267 (e).** If $A$ is closed, by proposition 267 (d),
\[
A \cup \text{derived}(A) = \text{cl}(A) = A,
\]
hence derived$(A) \subseteq A$.

**Proof of necessity of ??**. Dual to proposition 267 (a).

**Proof of 267 (c).** Consider the complement of derived$(A)$. If it is empty, derived$(A)$ is trivially closed. Otherwise, let $x \in X \setminus \text{derived}(A)$.

- If $x$ is an isolated point of $A$, by proposition 267 (b) there exists a neighborhood of $x$ that does not intersect $A \setminus \{x\}$.

- If $x$ is not a point of $A$, aiming at a contradiction, assume that every neighborhood of $x$ intersects $A$. Then, by definition 264 (b), $x \in \text{fr}(A)$. But fr$(A) \subseteq \text{cl}(A)$ and cl$(A) = \text{cl}(A \setminus \{x\})$ because $x$ does not belong to $A$. Therefore, $x$ is a cluster point of $A$. This contradicts our assumption that $x \notin \text{derived}(A)$, hence we can conclude that there exists a neighborhood of $X$ that does not intersect $A = A \setminus \{x\}$.

In both cases, proposition 245 allows us to conclude that $X \setminus \text{derived}(A)$ is open and, hence, derived$(A)$ is closed.

**Proof of 267 (d).** Clearly $A \subseteq \text{cl}(A)$. Also
\[
\text{derived}(A) \subseteq \bigcup_{x \in X} \text{cl}(A \setminus \{x\}) \subseteq \text{cl}(A).
\]

Now we will prove the reverse inclusion. Let $x \in \text{cl}(A)$. Then either $x \in A$ or $x \in \text{fr}(A)$. Assume the latter. By definition 264 (b), every neighborhood $U$ of $x$ has points both in $A$ and outside $A$, therefore $U \cap (A \setminus \{x\})$ is nonempty. By proposition 258 (b), $x \in \text{cl}(A \setminus \{x\})$, that is, $x \in \text{derived}(A)$.

**Proof of 267 (e).** Assume that derived$(A) \subseteq A$ and, aiming at a contradiction, suppose that $A$ is not closed. Fix a point $x \in \text{cl}(A) \setminus A$. By proposition 267 (d), this is a cluster point. By proposition 267 (a), every for neighborhood $U$ of $x$ the intersection $U \cap (A \setminus \{x\}) \subseteq U \cap A$ is nonempty. Since this holds for arbitrary neighborhoods, by proposition 258 (b), $A$ is closed.
Proof of sufficiency of ??: Special case of proposition 267 (e).

Proof of necessity of ??: We already know from proposition 267 (e) that it is sufficient for derived(A) to belong to A for A to be closed. But A \ derived(A) consists of all isolated points, therefore every point in A is either a cluster point or an isolated point.

Definition 268. Let (X, T) be a topological space and A ⊆ X be any set. We say that A is

(a) dense in X if cl A = X (if X is assumed from the context, we simply say that A is dense).

(b) codense in X if X \ A is dense, i.e. cl(X \ A) = X.

(c) nowhere dense in X if cl(A) is codense, i.e. X = cl(X \ cl A) = cl(int(X \ A)).

(d) dense in itself if A ⊆ derived(A), i.e. if A has no isolated points.

We define the density d(X) of X to be the minimum cardinality of all dense sets. If d(X) ≤ ℵ₀, we say that the space is separable.

Proposition 269. Dense sets have the following basic properties:

(a) The set A is dense if and only if every nonempty open set intersects A.

Proof.

Proof of 269 (a). Special case of proposition 258 (b).

Proposition 270. Nowhere dense sets have the following basic properties:

(a) Nowhere dense sets have an empty interior

(b) Nowhere dense sets are entirely contained in their boundaries.

(c) The set A is nowhere dense if and only if int(cl(A)) = ∅.

(d) The set is nowhere dense if and only if its closure does not contain a nonempty open set.

(e) The set A is nowhere dense if and only if every open set contains a nonempty open subset disjoint from A.

(f) A subset of a nowhere dense set is nowhere dense.

(g) The homeomorphic image of a nowhere dense set is nowhere dense.

(h) A set is closed and nowhere dense if and only if its complement is open and dense.

Proof.

Proof of 270 (c). Follows directly from proposition 262.

Proof of 270 (a). Follows from proposition 270 (c) because int(A) ⊆ int(cl(A)) = ∅.

Proof of 270 (b). Follows from proposition 270 (a) and definition 264 (a).
**Proof of 270 (d).** By proposition 269 (a), \(A\) is nowhere dense if and only if every nonempty open set intersects \(X \setminus \text{cl}(A) = \text{int}(X \setminus A)\). By proposition 245, the last condition is equivalent to every nonempty open set having a nonempty open subset in \(\text{int}(X \setminus A) = X \setminus \text{cl}(A)\), which in turn implies proposition 270 (d).

**Proof of 270 (f).** Let \(A\) be a nowhere dense set and let \(B \subseteq A\). Then
\[
\text{int}(\text{cl}(B)) \subseteq \text{int}(\text{cl}(A)) \subseteq \emptyset,
\]
therefore \(B\) is also nowhere dense.

**Proof of 270 (g).** Let \(f : X \to Y\) be a homeomorphic embedding (not necessarily surjective) and let \(A \subseteq X\) be a nowhere dense set. Let \(V\) be an open set in \(Y\). Then \(f^{-1}(V)\) is open in \(X\) and, by proposition 270 (e), there exists an open subset \(U \subseteq f^{-1}(V)\) that is disjoint from \(A\). Therefore, \(f(U) \subseteq f(f^{-1}(V)) \subseteq V\). Furthermore, \(f(U)\) is open and \(f(U) \cap f(A) = f(U \cap A) = f(\emptyset) = \emptyset\), therefore \(f(A)\) is nowhere dense.

**Proof of 270 (h).** If \(A\) is an open dense set, then \(X \setminus A\) is closed and
\[
\text{cl}(X \setminus \text{cl}(X \setminus A)) = \text{cl}(X \setminus (X \setminus A)) = \text{cl}(A) = X,
\]
therefore \(X \setminus A\) is nowhere dense.

**Definition 271.** Fix a topological space \(X\) and \(\mathcal{F} \subseteq \text{pow}(X)\). Denote by \(\mathcal{F}_\delta\) the family of all countable intersections of elements of \(\mathcal{F}\) and by \(\mathcal{F}_\sigma\) the family of all countable unions of elements of \(\mathcal{F}\).

The family \(\mathcal{F}_\delta\) is the family of countable unions of closed sets and \(G_\sigma\) is the family of countable intersections of open sets.
6.2. Topological nets

In this section, $X$ will denote an arbitrary nonempty topological space.

**Definition 272.** A net or generalized sequence or Moore-Smith sequence in a nonempty set $S$ is a family of elements of $S$ indexed by a nonempty directed set, i.e. a function from a nonempty directed set $(\mathcal{K}, \leq)$ to $S$. We use the conventional notation for indexed families:

$$\{x_k\}_{k \in \mathcal{K}},$$

because the preorder on the domain $\mathcal{K}$ is usually clear from the context.

If we know that the net is a sequence, we will usually use the notation for sequences given in definition 951.

Note that this definition does not actually require a topology on $S$. Some other important definitions also do not require topologies:

(a) We say that $\{x_k\}_{k \in \mathcal{K}}$ is frequently in the set $A \subseteq S$ if for every index $k_0 \in \mathcal{K}$ there exists an index $k \geq k_0$ such that $x_k \in A$.

(b) We say that $\{x_k\}_{k \in \mathcal{K}}$ is eventually in the set $A \subseteq S$ if there exists an index $k_0$ such that $x_k \in A$ whenever $k \geq k_0$. This is obviously a stronger condition.

(c) We say that the net $\{y_m\}_{m \in M} \subseteq S$ is a subnet of $\{x_k\}_{k \in \mathcal{X}} \subseteq S$ if there exists an embedding function $\varphi : M \rightarrow \mathcal{K}$ such that

(i) To every $k \in \mathcal{K}$ there corresponds $m \in M$ such that $\varphi(m) \geq k$.

(ii) For every $m \in M$ we have $x_{\varphi(m)} = y_m$.

**Proposition 273.** Nets have the following basic properties:

(a) “Eventually in” implies “frequently in”.

(b) The net $\{x_k\}_{k \in \mathcal{X}} \subseteq S$ is eventually in $A \subseteq S$ if and only if it is not frequently in $S \setminus A$.

**Proof.**

**Proof of 273 (a).** Suppose that the net $\{x_k\}_{k \in \mathcal{X}} \subseteq S$ is eventually in $A \subseteq S$. Then there exists an index $k_0$ such that $x_k \in A$ for all $k \geq k_0$.

Given any index $k_1$, we choose $k_2$ such that $k_2 \geq k_0$ and $k_2 \geq k_1$ (this is possible by the definition of a directed set). Then $x_{k_2} \in A$ and $k_2$ satisfies the existence quantifier in definition 272 (a).

**Proof of 273 (b).** Suppose that $\{x_k\}_{k \in \mathcal{X}}$ is both eventually in $A$ and frequently in $S \setminus A$.

Since the net is eventually in $A$, we can fix an index $k_0$ such that $x_k \in A$ whenever $k \geq k_0$.

Since the net is frequently in $A$, we can fix an index $k_1 \geq k_0$ such that $k_1 \in S \setminus A$, which is a contradiction.

This proves that the two conditions are incompatible. 

**Definition 274.** Let $X$ be a topological space and $\{x_k\}_{k \in \mathcal{X}} \subseteq X$ be a net.
(a) If the net is frequently in every neighborhood of \( x_0 \in X \), we say that \( x_0 \) is a **cluster point** or an **accumulation point** of \( \{x_k\}_{k \in \mathbb{X}} \subseteq X \).

(b) If the net is eventually in every neighborhood of \( x_0 \in X \), we say that \( x_0 \) is a **limit point** of \( \{x_k\}_{k \in \mathbb{X}} \subseteq X \).

In general, there can exist multiple limit points (see example 275) and even more cluster points (see example 276). In Hausdorff spaces, however, limits are unique by proposition 308. If \( \{x_k\}_{k \in \mathbb{X}} \subseteq X \) has a unique limit, we say that the net **converges**\(^2\) to \( x_0 \) use the notation

\[
x_0 = \lim_{k \in \mathbb{X}} x_k.
\]

If the net is a sequence, we also use the following notations:

- \( x_0 = \lim_{k \to \infty} x_k \)
- \( x_0 = \lim x_k \)
- \( x_k \xrightarrow[k \to \infty]{\text{}} x_0 \)
- \( x_k \to x_0 \)

**Example 275.** Even limits of sequences need not be unique in arbitrary topological spaces. Let \( X = \{y, z\} \) be a binary set with the indiscrete topology \( \{\emptyset, X\} \). Let

Define the following sequence

\[
x_k := \begin{cases} y, & \text{k is even,} \\ z, & \text{k is odd.} \end{cases}
\]

The only neighborhood of \( y \), the whole space \( X \), contains all members of the sequence, therefore \( y \) is a limit point of the sequence. The same is true for \( z \), however.

**Example 276.** Consider the net \( \{\sin(k)\}_{k \in \mathbb{R}} \). It has no limit point, yet every real number in the interval \([-1, 1]\) is a cluster point.

**Example 277.** A commonly used technique is to use a variation of a **reverse set inclusion net.**

Fix an element \( x_0 \in X \) of any topological space and choose an element \( x_U \) of every neighborhood \( U \) of \( x_0 \). Consider the directed set \( (\mathcal{T}(x), \subseteq) \) consisting of all neighborhoods of \( x_0 \) ordered by **reverse inclusion**, i.e. \( U \subseteq V \iff U \supseteq V \).

Choose an element \( x_U \) from each neighborhood \( U \) of \( x_0 \). Then, by construction, \( x_0 \) is a limit point of the net \( \{x_U\}_{U \in \mathcal{T}(x_0)} \).

**Proposition 278.** Convergence of nets has the following basic properties:

(a) The point \( x_0 \in X \) is a limit point of the sequence \( \{x_k\}_{k=1}^{\infty} \subseteq X \) if and only if, given a neighborhood \( U \) of \( x_0 \), only finitely many elements of the sequence are outside \( U \).

\(^2\)bg: схожда, ru: сходится

141
(b) Every limit point is a cluster point.

(c) A point $x_0 \in X$ is a cluster point of the net $\{x_k\}_{k \in X} \subseteq X$ if and only if $x_0$ is a limit point of some subnet.

(d) If a net has a limit point, all of its cluster points are limit points.

(e) A net has a unique limit point if and only if it has a unique cluster point.

(f) A net has a unique limit point if and only if all subnets have the same limit point.

Proof.

**Proof of 278 (a).** This is simply a restatement of definition 274 (b) for the special case of sequences.

**Proof of 278 (b).** Follows from proposition 273 (a).

**Proof of 278 (c).** The definition of a cluster point (definition 274 (a)) allows us to build a reverse inclusion net in the style of example 277.

**Proof of 273 (a).** Suppose that the net $\{x_k\}_{k \in X} \subseteq S$ is eventually in $A \subseteq S$. Then there exists an index $k_0$ such that $x_k \in A$ for all $k \geq k_0$.

Given any index $k_1$, we choose $k_2$ such that $k_2 \geq k_0$ and $k_2 \geq k_1$ (this is possible by the definition of a directed set). Then $x_{k_2} \in A$ and $k_2$ satisfies the existence quantifier in definition 272 (a).

**Proof of 273 (b).** Suppose that $\{x_k\}_{k \in X}$ is both eventually in $A$ and frequently in $S \setminus A$.

Since the net is eventually in $A$, we can fix an index $k_0$ such that $x_k \in A$ whenever $k \geq k_0$.

Since the net is frequently in $A$, we can fix an index $k_1 \geq k_0$ such that $k_1 \in S \setminus A$, which is a contradiction.

This proves that the two conditions are incompatible. \( \square \)

**Proposition 279.** Fix a set $A \subseteq X$. A point $x_0 \in X$ belongs to $\text{cl} \ A$ if and only if there exists a net $\{x_k\}_{k \in X} \subseteq A$ for which $x_0$ is a limit point.

By proposition 278 (c), we can consider cluster points of nets rather than limit points.

**Proof.** The complement of the empty set is the empty set, hence the statement of the proposition holds vacuously. Assume that $A$ is nonempty.

**Proof of sufficiency.** Suppose that $x_0 \in \text{cl} \ A$. If $x_0 \in A$, then the one-element net $(x_0)$ converges to $x_0$.

If $x_0 \in \text{fr} \ A$, by definition 264 (b), every neighborhood of $x_0$ contains points from $A$. Therefore, we can build reverse inclusion net in the style of example 277 that converges to $x_0$.

**Proof of necessity.** Let $x_0$ be a limit point of $\{x_k\}_{k \in X} \subseteq A$. We will show that $x_0$ belongs every closed set that contains $A$.

Let $F \supseteq A$ be a closed set. Denote $U := X \setminus F$. Suppose that $x_0 \in U$. Then $U$ is a neighborhood $x_0$ and, by definition 274 (a), the net $\{x_k\}_{k \in X} \subseteq A$ is eventually in $U$. But $U$ does not contains $A$.

The obtained contradiction shows that $x_0$ belongs to every closed set containing $A$ and hence to their intersection, the closure $\text{cl} \ A$. \( \square \)
Proposition 280. The point \( x_0 \in X \) is a cluster point of the set \( A \) if and only if it is a limit point of some net in \( A \setminus \{x_0\} \) (or, equivalently, a cluster point of some net in \( A \setminus \{x_0\} \)).

Proof.

Proof of sufficiency. Let \( x_0 \in \text{derived}(A) \). By proposition 267 (a), every neighborhood \( U \) of \( x_0 \) intersects \( A \setminus \{x_0\} \). Choose \( x_U \in U \cap (A \setminus \{x_0\}) \) for every neighborhood \( U \) of \( x_0 \) and form the reverse inclusion net \( \{x_U\}_{U \in T(x)} \). Then \( x_0 \) is a limit point of this net. Furthermore, the net is contained in \( A \setminus \{x_0\} \).

Proof of necessity. Conversely, if \( \{x_k\}_{k \in \mathbb{X}} \subseteq A \setminus \{x_0\} \) is a net and if \( x_0 \) is a limit point of this net, then for every neighborhood \( U \) of \( x_0 \) there exists an index \( k_U \) such that for \( k \geq k_U \) we have \( x_k \in U \). In particular, \( U \cap A \) contains elements other than \( x_0 \). Since this is true for any neighborhood \( U \) of \( x_0 \), by proposition 267 (a) we conclude that \( x_0 \) is a cluster point of the set \( A \).

Corollary 281. A set is closed if and only if it contains the limit points of all of its nets (or, equivalently, the cluster points of all of its nets).

Proof. By proposition 267 (e), the set \( A \) is closed if and only if it contains all of its cluster points. By proposition 280, this is equivalent to \( A \) containing all limit points of its nets.

Proposition 282. Fix a topological space \( X \), a point \( x_0 \) and a local subbase \( P(x_0) \). The point \( x_0 \) is a limit of the net \( \{x_k\}_{k \in \mathbb{X}} \subseteq X \) if and only if it is eventually in every element \( U_P \) of the local subbase \( P(x_0) \).

Proof.

Proof of sufficiency. Obvious consequence of the definition of local subbase.

Proof of necessity. Fix a neighborhood \( U \) of \( x_0 \). By definition 256, there exists a finite family \( \{U_k\}_{k=1}^n \subseteq P(x_0) \) such that \( \bigcap_{k=1}^n U_k \subseteq U \). Since the net \( \{x_k\}_{k \in \mathbb{X}} \subseteq X \) is eventually in each of \( U_k, k = 1, \ldots, n \), from transitivity of inclusion it follows that the net is eventually in \( U \).

Definition 283. In analogy to definition 257, we define the sequential closure operator

\[
\text{cl}^S : \text{pow}(X) \to \text{pow}(X)
\]

\[
\text{cl}^S(A) := \{ x \in X : x \text{ is a limit point of some sequence } \{x_k\}_{k=1}^\infty \subseteq A \}.
\]

If \( \text{cl}^S(A) = A \), we say that \( A \) is sequentially closed.

Definition 284. A topological space is called sequential if every sequentially closed set is closed.

Remark 285. By proposition 279, in a sequential space, a set is closed if and only if it is sequentially closed.

By corollary 281, a set is closed if and only if it contains the limit points of all of its nets.

Therefore, a set in a sequential space is closed if and only if it contains the limit points of all of its sequences.
Since we are able to define a topology in terms of closed sets, this means that the topology in a sequential space is completely determined by convergent sequences rather than convergent nets as in general topological spaces.

This allows us to restrict ourselves only to sequences rather than arbitrary nets in certain spaces like metric spaces.

**Lemma 286.** Let $X$ be a sequential space and $x_0$ limit point of the net $\{x_k\}_{k \in \mathcal{K}}$, then we can define a sequence

$$\{x_k\}_{k=1}^\infty \subseteq \{x_k : k \in \mathcal{K}\},$$

consisting of members of the net, for which $x_0$ is a limit point.

**Proof.** Let $X$ be a first-countable space and let $x_0$ be a limit point of the net $\{x_k\}_{k \in \mathcal{K}}$.

Since $X$ is a first countable space, we can fix a countable local base $\{U_k\}_{k=1}^\infty$ at $x_0$. For each $k = 1, 2, ..., $ define the neighborhood $V_k := \bigcap_{m=1}^k U_m$, so that $V_k \subseteq V_m$ whenever $k \geq m$.

For each neighborhood $V_k$, since $\{x_k\}_{k \in \mathcal{K}}$ is eventually in $V_k$, there exists an index $k_k$ such that $x_{k_k} \in V_k$.

Thus, we obtain a sequence $\{x_{k_k}\}_{k=1}^\infty$ that is eventually in every neighborhood of the local base $\{V_k\}_{k=1}^\infty$ of $x_0$, which by proposition 282 is sufficient for $x_0$ to be a limit point of the sequence. \qed

**Proposition 287.** Every first-countable space is sequential.

**Proof.** Let $X$ be a first-countable space and let $A \subseteq X$ be a sequentially closed set. We must show that it is closed.

Fix a point $x_0 \in \text{cl}(A)$. We will show that $x_0 \in A$. By proposition 279, there is a net $\{x_k\}_{k \in \mathcal{K}} \subseteq A$ for which $x_0$ is a limit point.

By lemma 286, we can choose a sequence $\{x_k\}_{k=1}^\infty$ that converges to $x_0$ out of elements of the net. But since $X$ is a sequential space, the limit points of any sequence are contained in the sequentially closed set $A$.

We showed that $A = \text{cl}(A)$. Since $A$ was an arbitrary sequentially closed set, we conclude that the space $X$ is sequential. \qed
6.3. Function convergence

**Definition 288.** Fix two topological spaces $X$ and $Y$. Let $A \subseteq X$ be a nonempty set and let $f : A \rightarrow Y$ be a function. We give two equivalent definitions for $y_0 \in Y$ being a **limit point** of $f$ at $x_0 \in \text{cl}(A)$. If $y_0$ is the unique limit point (e.g. in Hausdorff spaces), we write

$$\lim_{x \rightarrow x_0} f(x) = y_0.$$  

(a) (Cauchy-style condition) For every neighborhood $V$ of $y_0$ there exists a neighborhood $U$ of $x_0$ such that $f(U \cap A) \subseteq V$.

(b) (Heine-style condition) For every net $\{x_k\}_{k \in K} \subseteq A$, for which $x_0$ is a limit point, the corresponding net $\{f(x_k)\}_{k \in K}$ has $y_0$ as a limit point.

**Proof.**

**Proof that 288 (a) implies 288 (b).** Let $\{x_k\}_{k \in K} \subseteq U$ be a net with limit point $x_0$. Consider the net $\{f(x_k)\}_{k \in K}$. Fix a neighborhood $V$ of $y_0$. We need to show that $\{f(x_k)\}_{k \in K}$ is eventually in $V$.

By definition 288 (a), there exists a neighborhood $U$ of $x_0$ such that $f(U) \subseteq V$. Since $x_0$ is a limit point of $\{x_k\}_{k \in K}$, there exists an index $k_0$ such that for all $k \geq k_0$, $x_k \in U$ and therefore $f(x_k) \in V$. Hence, $\{f(x_k)\}_{k \in K}$ is eventually in $V$.

We conclude that $y_0$ is a limit point of the net $\{f(x_k)\}_{k \in K}$ and that the Heine-style condition is satisfied.

**Proof that 288 (b) implies 288 (a).** Suppose that definition 288 (b) holds while definition 288 (a) does not. Let $V$ be a neighborhood of $y_0$. Then there exists no neighborhood $U$ of $x_0$ such that $f(U) \subseteq V$.

For any neighborhood $U$ of $x_0$ and let $y_U \in f(U) \setminus V$ and $x_U \in f^{-1}(U)$, so that $f(x_U) = y_U$. Consider the families

$$\{x_U\}_{U \in T(x_0)}, \quad \{f(x_U)\}_{U \in T(x_0)},$$

ordered by reverse inclusion of the neighborhoods $T(x_0)$ of $x_0$.

Note that $x_0$ is a limit point of $\{x_U\}_{U \in T(x_0)}$. By definition 288 (b), $y_0$ is a limit point of $\{f(x_U)\}_{U \in T(x_0)}$. But this contradicts our choice of the nets because $f(x_U) \notin V$ for any $U \in T(x)$.

The obtained contradiction demonstrates that definition 288 (b) implies definition 288 (a).

**Proposition 289.** Fix two topological spaces $X$ and $Y$ and two points $x_0 \in X$ and $y_0 \in Y$. Let $P(x_0)$ and $P(y_0)$ be local subbases for the corresponding points. Then the function $f : X \rightarrow Y$ converges to $y_0$ at $x_0$ if and only if every $V_P \in P(y_0)$ there exists $U_P \in B(x_0)$ such that $f(U_P) \subseteq V_P$.

Compare this result to proposition 282.

**Proof.**
**Proof of sufficiency.** Obvious consequence of definition 288 (a).

**Proof of necessity.** Fix a neighborhood $V$ of $x$. We will show that definition 288 (a) holds.

Let $\{V_k\}_{k=1}^n \subseteq P(y_0)$ be a family such that $\bigcap_{k=1}^n V_k \subseteq V$ (such a family exists by definition of a local subbase). By the antecedent of the implication we are proving, for every $k = 1, \ldots, n$ there exists an $U_k \in P(x_0)$ such that $f(U_k) \subseteq V_k$. Then $U := \bigcap_{k=1}^n U_k$ is a neighborhood of $x_0$ and, furthermore,

$$f(U) = f\left(\bigcap_{k=1}^n U_k\right) \subseteq \bigcap_{k=1}^n f(U_k) \subseteq \bigcap_{k=1}^n V_k \subseteq V.$$ 

Therefore, definition 288 (a) holds. \qed
6.4. Topological continuity

**Definition 290.** We say that the function \( f : X \to Y \) between topological spaces is **continuous** at the point \( x_0 \in X \) if \( f(x_0) \) is a limit point of \( f \) at \( x_0 \).

If limit point is unique (e.g. in Hausdorff spaces), this condition can be formulated by “interchanging” \( \lim \) and \( f \) as follows:

\[
f(x_0) = f \left( \lim_{x \to x_0} x \right) = \lim_{x \to x_0} f(x).
\]

**Definition 291.** We say that the function \( f : X \to Y \) between topological spaces is **everywhere continuous** or simply **continuous** if and of the following conditions hold:

(a) \( f \) is continuous at every point of \( X \) in the sense of **definition 290**.

(b) For every open set \( V \in T \), the preimage \( f^{-1}(V) \) is open.

(c) For every closed set \( F \in F_T \), the preimage \( f^{-1}(F) \) is closed.

(d) There exists a base \( B_{T_Y} \subseteq T_Y \), such that for every \( V \in B_{T_Y} \), the preimage \( f^{-1}(V) \) is open.

(e) There exists a subbase \( P_{T_Y} \subseteq T_Y \), such that for every \( V \in P_{T_Y} \), the preimage \( f^{-1}(V) \) is open.

(f) For every set \( A \subseteq X \), \( f(\text{cl}(A)) \subseteq \text{cl}(f(A)) \).

We denote the set of all continuous functions from \( X \) to \( Y \) by \( C(X, Y) \).

**Proof.**

**Proof that 291 (a) implies 291 (b).** Follows from **definition 288 (a)**.

**Proof that 291 (b) implies 291 (c).** If \( F \in F_{T_Y} \) is a closed set, \( Y \setminus F \) is open, therefore \( f^{-1}(Y \setminus F) = X \setminus f^{-1}(F) \) is also open. Hence, \( f^{-1}(F) \) is closed.

**Proof that 291 (b) implies 291 (d).** \( F \) is a base of itself.

**Proof that 291 (d) implies 291 (e).** Every base is also a subbase.

**Proof that 291 (e) implies 291 (a).** Follows from the equivalences in **definition 288**.

**Proof that 291 (c) implies 291 (f).** Note that

\[
\text{976 (a)} A \subseteq f^{-1}(f(A)) \subseteq f^{-1}(\text{cl}(f(A))).
\]

Apply \( f \circ \text{cl} \) to the above chain of inclusions to obtain

\[
f(\text{cl}(A)) \subseteq f(\text{cl}(f^{-1}(\text{cl}(f(A))))) \subseteq \text{cl}(f(A)),
\]

which proves the implication.
Proof that 291 (f) implies 291 (c). Fix a closed set $F \subseteq Y$. Then

$$f(\text{cl}(f^{-1}(F))) \subseteq \text{cl}(f(f^{-1}(F))) \subseteq \text{cl}(F) = F.$$  \hspace{1cm} (108)

Since cl is monotone, we have

$$f(\text{cl}(f^{-1}(F))) \supseteq f(f^{-1}(F)) \supseteq F.$$  \hspace{1cm} (109)

From (108) and (109) it follows that

$$F = f(\text{cl}(f^{-1}(F))).$$

By taking the preimage, we obtain

$$f^{-1}(F) = f^{-1}(f(\text{cl}(f^{-1}(F)))) \supseteq \text{cl}(f^{-1}(F)).$$  \hspace{1cm} (b)

Therefore, $f^{-1}(F)$ is closed. \hfill \Box

Definition 292. We say that the continuous function $f : X \to Y$ is **open** (resp. **closed**), if the image $f(U)$ of an open (resp. closed) in $\mathcal{T}_X$ set is open (resp. closed) in $\mathcal{T}_Y$.

If $f$ is an open bijection, we say that $f$ is a **homeomorphism**. If $f$ is only an open injection, we say that $f$ is a **homeomorphic embedding**.

Definition 293. Let $I$ be an interval (of any type) in $\mathbb{R}$ with endpoints $a < b$, not necessarily finite. Depending on the use case, we define a **parametric curve** on $I$ by any of the non-equivalent definitions

(a) A continuous function $\gamma : I \to X$ is called a parametric curve.

(b) The image $\text{img}(\gamma)$ of a parametric curve $\gamma$ is also called a parametric curve.

(c) The equivalence class of all continuous functions from $I$ to $X$ with

$$\gamma \cong \beta \iff \text{img}(\gamma) = \text{img}(\beta)$$

and the endpoints of $\gamma$ and $\beta$ coincide

is also called a parametric curve.

The points $\gamma(a)$ and $\gamma(b)$ are called the **endpoints** of the curve, $\gamma(a)$ is the **start** and $\gamma(b)$ is the **end**. We say that $\gamma$ connects $a$ and $b$.

Parametric curves on $I = [0, 1]$ are also called **paths**.

We define some fundamental types of curves:

(a) The curve $\gamma$ is called **closed** if its endpoints coincide, i.e. $\gamma(a) = \gamma(b)$.

(b) The curve $\gamma$ is called **simple** if the function $\gamma : I \to Y$ is injective with the possible exception of the endpoints (in which case we speak of **simple closed curves**).

If $\gamma : I \to X$ is a parametric curve, related curves are:
(a) The graph $\text{gph}(\gamma)$ of $\gamma$ is a the image of the curve $\bar{\gamma}(t, x) := (t, \gamma(x))$ in the topological space $I \times X$.

(b) If $M$ is a subset of $X$ and if there exists a curve $\gamma : I \to X$ such that $\text{imag}(\gamma) = M$, we call $M$ an implicit parametric curve.

**Definition 294.** In analogy to definition 293 (and with the caveats of definition 293), we define parametric hypersurfaces as follows:

Let $\xi$ is a potentially infinite cardinal number, let $\text{card} \mathcal{K} = \xi$ and let $\{I_\alpha\}_{\alpha \in \mathcal{K}}$ be a family of intervals in $\mathbb{R}$. We define a parametric hypersurface to be a continuous image from the product space $\prod_{\alpha \in \mathcal{K}} I_\alpha$ to $Y$.

We call $\xi$ the dimension of the hypersurface.

**Definition 295.** TODO: Define fundamental groupoids.
6.5. Initial and final topologies

**Definition 296.** Since the topology $\mathcal{T}$ of a topological space $(X, \mathcal{T})$ consists of subsets of $X$, we cannot build a first-order theory from definition 243-definition 243. We can, however, explicitly describe the category $\textbf{Top}$ of topological spaces as

1116 (a) The class of objects is the class of all topological space.

1116 (b) The morphisms between two topological spaces are the continuous functions between them.

1116 (c) Composition of morphisms is the usual function composition.

**Theorem 297.** The category $\textbf{Top}$ of is both complete and cocomplete.

**Definition 298.** Let $(X_k, \mathcal{T}_k)_{k \in \mathcal{K}}$ be a family of topological spaces. Let $\mathcal{K}$ be a bare set and let

$$\{f_k : X \to X_k\}_{k \in \mathcal{K}}$$

be a family of functions.

The topology on $X$ generated by the subbase

$$\mathcal{P} := \{f_k^{-1}(U) : k \in \mathcal{K}, U \in T_k\}$$

is called the initial (or weak) topology on $X$ generated by the family $\{f_k\}_{k \in \mathcal{K}}$.

It is the weakest topology that makes all functions in the family $\{f_k\}_{k \in \mathcal{K}}$ continuous.

**Definition 299.** Dually, if the family of functions is of the type

$$\{f_k : X_k \to X\}_{k \in \mathcal{K}},$$

then we define the final (or strong) topology on $X$ generated by the family $\{f_k\}_{k \in \mathcal{K}}$ as the topology

$$\mathcal{J} := \{U \subseteq X : \forall k \in \mathcal{K}, f_k^{-1}(U) \in T_k\}.$$ 

It is the strongest topology that makes all functions in the family $\{f_k\}_{k \in \mathcal{K}}$ continuous.

**Proposition 300.** Let $D : I \to \textbf{Top}$ be a small diagram. For each space in the image $D(I)$, denote the set corresponding by $X_k$ and the corresponding topology by $\mathcal{T}_k$.

The limit (resp. colimit) $(X, \mathcal{T})$ of $D$ can then be described as

(a) $(X, \{f_k\}_{k \in I}) = \lim UD$ (resp. $\lim UD$) is the limit (resp. colimit) in $\textbf{Set}$ of $U \circ D$, where $U : \textbf{Top} \to \textbf{Set}$ is the forgetful functor.

(b) $\mathcal{T}$ is the initial (resp. final) topology on $X$ generated by the family of functions $\{f_k\}_{k \in I}$.

**Definition 301.** Let $(X, \mathcal{T})$ be a topological space and let $M \subseteq X$ be a subset of $X$. The **topological subspace** $(M, \mathcal{T}_M)$ is obtained by endowing $M$ with the topology

$$\mathcal{T}_M := \{U \cap M : U \in \mathcal{T}\}.$$ 

The topology $\mathcal{T}_M$ is called the **subspace topology** or **induced topology**. It is the initial topology generated by the canonical embedding $\iota : M \to X$.  

150
Definition 302. The **topological product** or **Tychonoff product**

\[
\left( \prod_{k \in \mathcal{K}} X_k, \prod_{k \in \mathcal{K}} \mathcal{T}_k \right)
\]

of the family \((X_k, \mathcal{T}_k)_{k \in \mathcal{K}}\) is simply the categorical product in the category \textbf{Top} (see definition 1213). The underlying set \(\prod_{k \in \mathcal{K}} X_k\) is the Cartesian product and the topology \(\prod_{k \in \mathcal{K}} \mathcal{T}_k\) is called the **product topology**.

Let \((X_k, \mathcal{T}_k)_{k \in \mathcal{K}}\) and \((Y_k, \mathcal{O}_k)_{k \in \mathcal{K}}\) be two families of topological spaces and let

\[
\{f_k : X_k \to Y_k\}_{k \in \mathcal{K}}
\]

be a family of arbitrary functions between them.

We define the **product** \(\prod_{k \in \mathcal{K}} f_k\) of \(\{f_k\}_{k \in \mathcal{K}}\) as the function

\[
\left( \prod_{k \in \mathcal{K}} f_k \right) : \prod_{k \in \mathcal{K}} X_k \to \prod_{k \in \mathcal{K}} Y_k
\]

\[
\left( \prod_{k \in \mathcal{K}} f_k \right) \{x_k\}_{k \in \mathcal{K}} := \{f_k(x_k)\}_{k \in \mathcal{K}}.
\]

If all of the spaces \((X_k, \mathcal{T}_k)\) are equal to some space \((X, \mathcal{T})\), we call the product of \(\{f_k\}_{k \in \mathcal{K}}\) the **diagonal product** and denote it by

\[
\Delta_{k \in \mathcal{K}} f_k : X \to \prod_{k \in \mathcal{K}} Y_k.
\]

Definition 303. Let \(X\) be a topological space and let \(\equiv\) be an equivalence relation on \(X\). The **quotient space** \((X, \mathcal{T}) / \sim\) is obtained by endowing the quotient set \(X / \equiv\) with the final topology given by the canonical projection map \(x \mapsto [x]\).

Definition 304. The **topological direct sum**

\((\oplus_{k \in \mathcal{K}} X_k, \oplus_{k \in \mathcal{K}} \mathcal{T}_k)\)

of the family \((X_k, \mathcal{T}_k)_{k \in \mathcal{K}}\) is simply the categorical coproduct in the category \textbf{Top} (see definition 1213). The underlying set \(\oplus_{k \in \mathcal{K}} X_k\) is the disjoint union and the topology \(\oplus_{k \in \mathcal{K}} \mathcal{T}_k\) is called the **direct sum topology**.

Let \((X_k, \mathcal{T}_k)_{k \in \mathcal{K}}\) and \((Y_k, \mathcal{O}_k)_{k \in \mathcal{K}}\) be two families of topological spaces and let

\[
\{f_k : X_k \to Y_k\}_{k \in \mathcal{K}}
\]

be a family of arbitrary functions between them. Let \(\iota_{X_k} : X_k \to \oplus_{k \in \mathcal{K}} X_k\) and \(\iota_{Y_k} : Y_k \to \oplus_{k \in \mathcal{K}} Y_k\) be the corresponding canonical embeddings.

We define the **direct sum** \(\oplus_{k \in \mathcal{K}} f_k\) of \(\{f_k\}_{k \in \mathcal{K}}\) as the function

\[
(\oplus_{k \in \mathcal{K}} f_k) : \oplus_{k \in \mathcal{K}} X_k \to \oplus_{k \in \mathcal{K}} Y_k
\]
\[(\oplus_{k \in \mathcal{K}} f_k)|_{X_k} := t_{Y_k} \circ f_k.\]

Obviously \(\oplus_{k \in \mathcal{K}} f_k\) is continuous whenever all \(f_k\) are continuous.

If all of the spaces \((Y_k, \mathcal{O}_k)\) are equal to some space \((Y, \mathcal{O})\), we call the direct sum of \(\{f_k\}_{k \in \mathcal{K}}\) simply a sum and denote it by

\[\sum_{k \in \mathcal{K}} f_k : \oplus_{k \in \mathcal{K}} X_k \to Y.\]
6.6. Separation axioms

**Definition 305.** Two subsets $A, B \subseteq X$ of a topological space $(X, \mathcal{T})$ are called **separated** or **separated using neighborhoods** if there exist disjoint open sets $U \supseteq A$ and $V \supseteq B$. In particular, two points are separated if their respective singleton sets are separated.

We say that $A$ and $B$ are **functionally separated** if there exists a continuous function $f : X \to [0, 1]$ such that $f(A) = 0$ and $f(B) = 1$.

**Definition 306.** We can classify topological spaces using the following separation axioms. Fix a topological space $(X, \mathcal{T})$.

- **$T_0$** (Kolmogorov) $X$ is $T_0$ if for every two different points $x, y \in X$, there exists an open set $U \in \mathcal{T}$ such that either $x \in U$ or $y \in U$.

- **$T_{0.5}$** $X$ is $T_{0.5}$ if every singleton set $\{x\}$ is either open or closed.

- **$T_1$** (Frechet) $X$ is $T_1$ if every singleton set $\{x\}$ is closed.

- **$T_2$** (Hausdorff) $X$ is $T_2$ if every two different points $x, y \in X$ can be separated using neighborhoods, i.e. there exist disjoint open sets $U \ni x$ and $V \ni y$.

- **$T_3$** $X$ is **regular** if every point and every closed set can be separated using neighborhoods.

  If in addition to being regular $X$ is $T_0$, we say that $X$ is a $T_3$ space.

- **$T_{3.5}$** (Tychonoff) $X$ is **completely regular** if every point and every closed set can be functionally separated.

  If in addition to being completely regular $X$ is $T_0$, we say that $X$ is a $T_{3.5}$ space.

- **$T_4$** (Urysohn) $X$ is **normal** if every two closed sets $F, G \in \mathcal{F}_X$ can be separated using neighborhoods, i.e. there exist disjoint open sets $U \supseteq F$ and $V \supseteq G$.

  If in addition to being normal $X$ is $T_1$, we say that $X$ is a $T_4$ space.

- **$T_5$** If every subspace of a $T_4$ space $X$ is $T_4$, we say that $X$ is a $T_5$ space or a **completely normal space**.

- **$T_6$** If every closed set in a $T_4$ space $X$ is $G_δ$ (see definition 271), we say that $X$ is a $T_6$ space or a **perfectly normal space**.

**Proposition 307.** Each numbered axiom in definition 306 implies the previous one.

**Proposition 308.** A topological space is Hausdorff if and only if every net has at most one limit.

**Proof.**

**Proof of sufficiency.** Let $X$ be Hausdorff and assume that there exists a net $\{x_k\}_{k \in \mathcal{K}}$ such that $y$ and $z$ are not necessarily distinct limit points.

Fix neighborhoods $U$ of $y$ and $V$ of $z$. Since both are limit points, there exist indices $k_U$ and $k_V$ such that $k \geq k_U$ implies $x_k \in U$ and $i \geq i_k$ implies $x_i \in V$. 

153
Since $\mathcal{K}$ is a directed set, there exists an upper bound $k_0$ of $k_U$ and $k_V$. Thus,

$$x_k \in U \cap V \quad \forall k \geq k_0.$$ 

In particular, the intersection $U \cap V$ is nonempty and is a neighborhood of both $y$ and $z$. If $y \neq z$, then we have two distinct points such that no two neighborhoods of $y$ and $z$, respectively, are disjoint. This contradicts the assumption that $X$ is Hausdorff. Thus, $y = z$.

**Proof of necessity.** Conversely, if $X$ is not Hausdorff, then for every two distinct points $y$ and $z$ and every two neighborhoods $U \ni y$ and $V \ni z$, their intersection $U \cap V$ is nonempty.

Let $\mathcal{U}$ and $\mathcal{V}$ be the sets of all neighborhoods of $y$ and $z$, respectively. Since they are both partially ordered by set inclusion $\subseteq$, define the directed set $(\mathcal{U} \times \mathcal{V}, \leq)$ with order

$$(U, V) \leq (U', V') \iff U \supset V \text{ and } U' \supset V'.$$

For each $(U, V) \in \mathcal{U} \times \mathcal{V}$, choose a point $x_{(U, V)}$ from $U \cap V$.

Thus, the net $\{x_{(U, V)}\}_{(U, V) \in \mathcal{U} \times \mathcal{V}}$ has both $y$ and $z$ as its limit points, which contradicts our initial assumption.  

**Lemma 309** (Urysohn’s lemma). In a normal space, every pair $A, B$ of disjoint closed sets can be functionally separated.

**Theorem 310.** Fix is an indexed family $\{X_k\}_{k \in \mathcal{K}}$ of topological spaces. Denote their product by $X$.

(a) [Eng89, theorem 2.3.11] If each one of $X_k$ is a $T_i$ space for $T_0$-T3.5, then $X$ is also a $T_i$ space.

(b) [Eng89, theorem 2.3.11] If $X$ is a $T_i$ space for $T_0$-T6, then each component $X_k$ is also a $T_i$ space.
6.7. Connected spaces

**Definition 311.** We say that the topological space \( X \) is **connected** if it satisfies any of the following equivalent conditions:

(a) If \( X = X_1 \cup X_2 \) and \( X_1, X_2 \) are disjoint open sets, either \( X_1 \) or \( X_2 \) is empty.

(b) If \( X = X_1 \cup X_2 \) and \( X_1, X_2 \) are disjoint closed sets, either \( X_1 \) or \( X_2 \) is empty.

(c) If \( X = X_1 \cup X_2 \) and \( X_1, X_2 \) are separated, either \( X_1 \) or \( X_2 \) is empty.

(d) The only subsets of \( X \) that are both open and closed are \( \emptyset \) and \( X \).

(e) Every continuous mapping \( f : X \to \{0, 1\} \) into the two-point discrete space is constant.

**Definition 312.** We say that \( X \) is **locally connected** if for every point \( x \in X \) and every neighborhood \( U \) of \( x \) there exists a connected set \( C \subseteq U \) such that \( x \in \text{int}(C) \).

**Definition 313.** We say that a topological space is **path connected** if every two points can be connected via a path.

**Definition 314.** We say that \( X \) is **locally path connected** if for every point \( x \in X \) and every neighborhood \( U \) of \( x \) there exists a neighborhood \( V \) of \( x \) such that for any \( y \in V \) there exists a path \( \gamma : [0, 1] \to U \) connecting \( x \) with \( y \).

**Proposition 315.** If \( X \) is connected and \( f : X \to Y \) is a homeomorphism, then \( Y \) is also connected.

**Proof.** Let \( Y = Y_1 \cup Y_2 \), where \( Y_1 \) and \( Y_2 \) are disjoint and open.

Note that the preimages \( \gamma^{-1}(Y_1) \) and \( \gamma^{-1}(Y_2) \) are open and disjoint, hence \( X = \gamma^{-1}(Y_1) \cup \gamma^{-1}(Y_2) \). But \( X \) is connected and by definition 311 (a), either \( \gamma^{-1}(Y_1) \) or \( \gamma^{-1}(Y_2) \) is the null set. Thus, either \( Y_1 \) and \( Y_2 \) is the null set and, again, by definition 311 (a), \( Y \) is connected. \( \square \)

**Proposition 316.** Any path connected space is connected.

**Proof.** Let \( X = X_1 \cup X_2 \), where \( X_1 \) and \( X_2 \) are disjoint and open.

Assume that both are nonzero and take \( x_1 \in X_1, x_2 \in X_2 \). Then there exists a path \( \gamma : I \to X \) with endpoints \( x_1 \) and \( x_2 \). Note that the preimages \( \gamma^{-1}(X_1) \) and \( \gamma^{-1}(X_2) \) are nonempty and open, hence cannot be separated by definition 311 (c). But this contradicts the disjointedness of \( X_1 \) and \( X_2 \).

The obtained contradiction proves that \( X \) is connected. \( \square \)
### 6.8. Compact spaces

**Definition 317.** The nonempty family \( F \) of subsets of the topological space \( X \) is said to be a [centered family of sets](#) or to have the [finite intersection property](#) if the intersection \( F_1 \cap \cdots \cap F_n \) of any finite collection of sets is nonempty.

**Definition 318.** The space \( X \) is called [compact](#) if any of the following equivalent finiteness conditions hold:

(a) Every open cover of \( X \) has a finite subcover.

(b) Every centered family \( F \) of closed subsets of \( X \) has a nonempty intersection.

(c) Every net has a cluster point or, equivalently, a [convergent](#) subnet. This property is also called “generalized sequential compactness” or, when restricted to sequences instead of general nets, simply “sequential compactness”.

**Proof.**

**Proof that 318 (a) implies 318 (b).** Assume that every open cover of \( X \) has a finite subcover. Let \( F \) be a centered family of closed subsets of \( X \). Aiming at a contradiction, suppose that \( \bigcap F = \emptyset \). Then

\[
X = X \setminus \bigcap F = \bigcup_{F \in F} (X \setminus F),
\]

which has a finite subcover indexed by, say, \( F' \subseteq F \). But \( F \) is a centered family and \( \bigcap F' \) is nonempty, hence

\[
X = \bigcup_{F \in F'} (X \setminus F) = X \setminus \bigcap F' \neq X.
\]

The obtained contradiction shows that \( \bigcap F \) is nonempty.

**Proof that 318 (b) implies 318 (a).** Assume that every centered family of closed sets has a nonempty intersection. Let \( \{U_k\}_{k \in K} \) be an open cover of \( X \). By putting \( F_k := U_k \) for all \( k \in K \), we obtain a family \( \{F_k\}_{k \in K} \) of closed sets with an empty intersection. Therefore, it is not a centered family. Then there exists at least one finite subfamily \( \{F_k\}_{k \in K'} \) with an empty intersection. The complement of this subfamily is then a finite cover of \( X \), which proves our statement.

**Proof that 318 (a) implies 318 (c).** Assume that every open cover of \( X \) has a finite subcover. Fix a net \( \{x_k\}_{k \in K} \subseteq X \).

Aiming at a contradiction, suppose that the net has no cluster points. For any point \( x \in X \) and any neighborhood \( U_x \) of \( x \), the net is not frequently in \( U_x \). Obviously \( \{U_x\}_{x \in X} \) is an infinite open cover of \( X \). Then it has a finite subcover indexed by, say, \( X' \subseteq X \).

Therefore, every element of the net \( \{x_k\}_{k \in K} \) if contained in one of the finitely many neighborhoods \( \{U_x\}_{x \in X'} \), and the net itself is frequently in at least one of the neighborhoods.

Thus, one of the finitely many points in \( X' \) is a cluster point of \( \{x_k\}_{k \in K} \).

---

156
Theorem 321 (Tychonoff’s product theorem). Let \((X_k, \mathcal{T}_k)_{k \in \mathcal{K}}\) be a family of topological spaces. Their product \(\prod_{k \in \mathcal{K}} X_k, \prod_{k \in \mathcal{K}} \mathcal{T}_k\) is compact if and only if \((X_k, \mathcal{T}_k)\) is compact for every \(k \in \mathcal{K}\).

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (n).

Theorem 322 (Weierstrass’ extreme value theorem). Let \(X\) be a compact topological space and let \(f : X \to \mathbb{R}\) be a continuous function into the real numbers.

Then \(f\) is bounded and there exist \(m, M \in X\) such that

\[
    f(m) = \min_{x \in X} f(x) \quad \text{and} \quad f(M) = \max_{x \in X} f(x).
\]

Definition 323. A topological space is called \textbf{locally compact} if every point has a relatively compact neighborhood.
6.9. Baire spaces

Remark 324. René-Louis Baire introduced the concept of Baire categories in 1899, almost 50 years before Samuel Eilenberg and Saunders MacLane introduced categories in [EM45] (see section 14 (Category theory) for the latter).

Unfortunately, topology utilizes both concepts, so the word “category” should be used with caution. To circumvent this, we use alternative terminology for Baire categories.

Definition 325. Any countable union of nowhere dense sets is called meager or a first category set (see remark 324 for terminology). If a set is not meager, we call it nonmeager or a second category set.

Proposition 326. Meager sets have the following basic properties (compare to proposition 270):

(a) A countable union of meager sets is meager.

(b) A subset of a meager set is meager.

(c) The homeomorphic image of a set A is meager if and only if A itself is meager.

Proof.

Proof of 326 (a). Follows from proposition 1065.

Proof of 326 (b). Fix a meager set A and let $B \subseteq A$. Then $A = \bigcup_{k=1}^{\infty} A_k$ for some nowhere dense sets $A_1, A_2, \ldots$. By proposition 270 (f), the sets $A_1 \cap B, A_2 \cap B, \ldots$ are also nowhere dense. But

$$B = A \cap B = \left( \bigcup_{k=1}^{\infty} A_k \right) \cap B = \bigcup_{k=1}^{\infty} (A_k \cap B).$$

Therefore, B is also nowhere dense.

Proof of 326 (c).

Proof of necessity. If A is meager, any homeomorphic image of A is meager by proposition 977 (b) and proposition 270 (g).

Proof of sufficiency. If $f : X \rightarrow Y$ is a homeomorphism and $f(A)$ is meager for some $A \subseteq X$, then $A$ is the homeomorphic image of the meager set $f(A)$ under $f^{-1}$ and is thus meager.

Definition 327. A topological space is called a Baire space if any of the following equivalent conditions hold:

(a) Every nonempty open set is nonmeager.

(b) A countable intersection of dense sets is dense.

Proof.

Proof of equivalence of 327 (a) and 327 (b). Follows from proposition 270 (h) and theorem 1279 (De Morgan’s laws).
**Proposition 328.** Every open subspace of a Baire space is a Baire space.

*Proof.* Let \((X, \mathcal{T})\) be a Baire space and let \((X', \mathcal{T}'_{X'})\) be an open subspace with the canonical embedding \(\iota : X' \to X\). The proposition holds vacuously if \(X' = \emptyset\), so we assume that \(X' \neq \emptyset\).

Note that \(\iota\) is continuous by definition, however it is also an open map because if \(U \in \mathcal{T}_{X'}\), then \(\iota(U) = U \cap X'\) is open in \(X\) as the intersection of two open sets. Therefore, it is a homeomorphic embedding and by **proposition 326 (c)**, \(U\) is meager if and only if \(\iota(U)\) is meager. Since \(X\) is a Baire space, \(\iota(U)\) is not meager and hence \(U\) is also not meager.

We showed that every nonempty open set \(U \in \mathcal{T}_{X'}\) is nonmeager, therefore \(X'\) is a Baire space. \(\square\)

**Theorem 329** (Baire category theorem).

(a) Complete metric spaces are Baire spaces.

(b) Locally compact Hausdorff spaces are Baire spaces.
6.10. Uniform spaces

**Remark 330.** Uniform spaces are an extension of both metric spaces and topological groups (including topological vector spaces). They are topological spaces that are “uniform” in the sense that different parts of the space behave the same, unlike manifolds.

In metric spaces, we use the notation \( \mu(x, y) < \varepsilon \) to mean that \( x \) and \( y \) are close (at a distance less than \( \varepsilon \)).

In (additive) topological groups, we instead have linear operations and use \( x - y \in U \) to mean that \( x \) and \( y \) are close (their difference belongs to some neighborhood of 0).

A proper generalization needs to make both metric spaces and topological groups feel natural as special cases. Generalizing metric space balls or neighborhoods of zero are nice options which unfortunately introduces some asymmetry since, for example for metric spaces, \( \mu(x, y) < \varepsilon \) can be written as either \( y \in B(x, \varepsilon) \) or \( x \in B(y, \varepsilon) \). This approach does not go far beyond what general topological spaces offer as a notation.

[Eng89, section 8] uses the notation \( |x - y| < V \) to mean that \( x \) and \( y \) belong to the entourage \( V \). This is a bit confusing because no absolute value nor subtraction are defined in uniform spaces. We find it simpler to not introduce any special notation beyond that of relations and, so we denote the same by \( (x, y) \in V \).

**Definition 331.** Let \( X \) be a set. For two binary relations \( V \) and \( U \) on \( X \) we define their sum as

\[
V + U := \{(x, z) : \exists y \in X : (x, y) \in U, (y, z) \in V\}
\]

and \( nV \) by \( n \)-fold iterative addition.

For any relation \( V \), we denote by \( -V \) the inverse relation.

A relation \( V \) on \( X \) is called an entourage if \( V \) is reflexive and symmetric.

In analogy to metric spaces, we define

(a) We define the open ball or simply ball with center \( x \) and radius \( V \) to be the set

\[
B(x, V) := \{y \in X : (y, x) \in V\}.
\]

(b) We say that the set \( S \subseteq X \) is bounded if it is contained in some ball.

**Proposition 332.** Using the notation of definition 331, we obtain properties similar to those of metrics:

\[
M1 \quad (x, x) \in V
\]

\[
M2 \quad (x, y) \in V \text{ if and only if } (y, x) \in V
\]

\[
M3 \quad \text{If } (x, y) \in U \text{ and } (y, z) \in V, \text{ then } (x, y) \in U + V.
\]

**Definition 333.** A uniform space is a set \( X \) with a nonempty family \( V \) of entourages on \( X \) such that

\[
U1 \quad \text{If } V \in V \text{ and } V \subseteq W \text{ for some entourage } W \text{ on } X, \text{ then } W \in V.
\]

\[
U2 \quad \text{If } V_1, V_2 \in V, \text{ then } V_1 \cap V_2 \in V.
\]
For every $V \in \mathcal{V}$ there exists $W \in \mathcal{V}$ such that $2W \subseteq V$.

$\mathcal{V} \cap \mathcal{V} = \Delta_X$, where $\Delta_X$ is the diagonal relation.

The family $\mathcal{V}$ is called a **uniform structure** or **uniformity** on $X$.

**Definition 334.** Let $(X, \mathcal{V})$ be a uniform space. We define its **uniform topology** or **induced topology** as the topology generated by the neighborhood system

$$\mathcal{B}(x) := \{B(x, V) : V \in \mathcal{V}\}.$$ 

If for some topological space $(X, \mathcal{I})$ there exists a uniformity such that $\mathcal{I}$ is its induced topology, we say that the topology $\mathcal{I}$ is **uniformizable**.

**Proof.** This proof of correctness does not actually rely on the uniform structure (except for $\mathcal{V}$ being nonempty), but rather on the properties of entourages.

It is indeed a neighborhood system because

**BP1** Every entourage is reflexive, hence $x$ is contained in every ball in $\mathcal{B}(x)$.

**BP2** For $B(x, U)$ and $B(x, V)$ we have

$$B(x, U \cap V) = \{y \in X : (x, y) \in U \cap V\} = \{y \in X : (x, y) \in U \text{ and } (x, y) \in V\} = B(x, U) \cap B(x, V).$$

**BP3** Fix $x, y \in X$ and a ball $B(y, V) \in \mathcal{B}(y)$ that contains $x$. We will show that $B(y, V) \subseteq B(x, 2V)$.

Fix $z \in B(y, V)$. We have $(z, y) \in V$. Then $(z, x) \in V + V = 2V$. Since $z \in B(y, V)$ was arbitrary, we conclude that $B(y, V) \subseteq B(x, 2V)$.

**Theorem 335.** A topological space is uniformizable if and only if it is a Tychonoff space.

**Definition 336.** Fix a uniform space $(X, \mathcal{V})$. The subfamily $\mathcal{B} \subseteq \mathcal{V}$ if entourages is called a **base** for $\mathcal{V}$ if every entourage $V \in \mathcal{V}$ contains a member of $\mathcal{B}$.

**Definition 337.** Let $X$ be an arbitrary set and let $\mathcal{B}$ be a family of entourages satisfying the following axioms:

**BU1** If $V_1, V_2 \in \mathcal{B}$, there exists an entourage $V \in \mathcal{B}$ such that $V \subseteq V_1 \cap V_2$.

**BU2** For every $V \in \mathcal{B}$ there exists $W \in \mathcal{B}$ such that $2W \subseteq V$.

**BU3** $\bigcap \mathcal{B} = \Delta_X$
Then the family of entourages
\[ \mathcal{V} := \{ V \subseteq X \times X : \exists B \in \mathcal{B} : B \in V \text{ and } V \text{ is reflexive and symmetric} \} \] (110)
is a uniform structure on \( X \). Furthermore, \( \mathcal{B} \) is a base of \( \mathcal{V} \).

In particular, the base on any topology satisfies definition 337 – definition 337.

**Lemma 338.** In a uniform space \((X, \mathcal{V})\), for every neighborhood \( A \) (in the topology) of a point \( x_0 \in X \) there exists an entourage \( V \in \mathcal{V} \) such that \( B(x_0, V) \subseteq A \).

**Proof.** By definition 334 and definition 247 (a), \( A \) is a union of balls centered at \( x_0 \). For any ball \( B(x_0, V) \) of this union, we have \( B(x_0, V) \subseteq A \).

**Proposition 339.** The uniform topology \( \mathcal{T} \) on \((X, \mathcal{V})\) the following basic properties:

(a) All balls are open sets.

(b) Every neighborhood of every point a ball centered at that point.

**Proof.**

**Proof of 339 (a).** We defined the balls to be the base of the uniform topology, therefore they are open.

339 (b) Fix a point \( x_0 \). It is a trivial consequence of definition 247 (b) that every neighborhood of \( x_0 \) contains some ball centered at a point that is not necessarily \( x_0 \). By proposition 254, this ball contains another ball centered at \( x_0 \).

**Proposition 340.** Fix a topological space \((X, \mathcal{T})\) and a uniform space \((Y, \mathcal{U})\). Let \( A \subseteq X \) be a nonempty set and let \( f : A \to Y \) be a function. Then \( y_0 \) is a limit point of \( f \) at \( x_0 \in X \) in the sense of definition 290 if and only if

\[ \forall V \in \mathcal{V} \exists A \in T(x_0) : x \in A \implies (f(x), y_0) \in V. \] (111)

If instead, \((X, \mathcal{U})\) is a uniform space, then \( y_0 \) is a limit point of \( f \) at \( x_0 \in X \) if and only if

\[ \forall V \in \mathcal{V} \exists U \in \mathcal{U} : (x, x_0) \in U \implies (f(x), y_0) \in V. \] (112)

Note that the limit point may not be unique because uniform spaces are not Hausdorff in general.

**Proof.** We will only prove eq. (112) because the proof of eq. (111) is a special case.

**Proof of sufficiency.** Suppose that \( y_0 \) is a limit point of \( f \) at \( x_0 \) and fix a neighborhood \( B \) of \( y_0 \). Then there exists a neighborhood \( A \) of \( x_0 \) such that \( f(A) \subseteq B \).

Fix an entourage \( V \in \mathcal{V} \). Then \( B(y_0, V) \) is also a neighborhood of \( y_0 \). By lemma 338 and definition 333, there exists an entourage \( V' \subseteq V \) such that \( B(f(x), V') \subseteq B \cap B(y_0, V) \).

Fix an entourage \( U \subseteq \mathcal{U} \) such that \( B(x_0, U) \subseteq A \). Then for any \( x \in X \), if \( (x, x_0) \in U \), we have \( (f(x), y_0) \in V' \). But \( V' \subseteq V \), therefore

\[ (x, x_0) \in U \implies (f(x), y_0) \in V. \]

This concludes the proof.
Proof of necessity. Fix a neighborhood $B$ of $y_0$ and an entourage $V \in V$ such that $B(x_0, V) \subseteq B$ (see lemma 338 for a justification). Then there exists $U \in U$ such that

$$(x, x_0) \in U \implies (f(x), y_0) \in V.$$  

Therefore, $A := B(x_0, U)$ is a neighborhood of $x_0$ such that $f(A) \subseteq B$. \hfill \qed

**Corollary 341.** A function $f : (X, V) \to (Y, U)$ between uniform spaces is continuous at $x_0 \in X$ if and only if

$$\forall V \in V \exists U \in U : (x, x_0) \in U \implies (f(x), f(x_0)) \in V.$$ 

**Definition 342.** Fix a set $X$ and a uniform space $(Y, V)$. Fix a function $f : X \to Y$.

(a) We say that the function $f : X \to Y$ is **bounded** if $f(X)$ is a bounded set, that is, if there exists a ball $B(y, V)$ such that $f(X) \subseteq B(y, V)$.

(b) We say that the family of functions $F$ from $X$ to $Y$ is **bounded** at $x_0$ if there exists a ball $B(y, V)$ such that the set $F(x_0) := \{f(x_0) : f \in F\}$ is contained in $B(y, V)$.

(c) We say that $F$ is **pointwise bounded** on the set $S \subseteq X$ if

$$\forall x \in S \exists B(y, V) : F(x) \subseteq B(y, V).$$

(d) We say that $F$ is **uniformly bounded** on $S \subseteq X$ if

$$\exists B(y, V) \forall x \in S : F(x) \subseteq B(y, V).$$

(e) If there is a topology $T$ on $X$, we say that the function $f : X \to Y$ is **locally bounded** if there exists an entourage $V \in V$ such that for each neighborhood $A \in T(x)$ we have $\text{diam } f(A) < V$.

**Proposition 343.** Let $(X, T)$ be a topological space and $(Y, V)$ be a uniform space. Any continuous function from $X$ to $Y$ is locally bounded.

**Proof.** Trivial. \hfill \qed

**Definition 344.** Fix a set $X$ and a uniform space $(Y, V)$. Let $\{f_k\}_{k \in K}$ be a net of functions from $X$ to $Y$. We say that $\{f_k\}_{k \in K}$ **converges pointwise** to the function $f$ and write $f_k \to f$ if

$$\forall V \in V \ \forall x \in X \ \exists k_0 \in K : k \geq k_0 \implies (f_k(x), f(x)) \in V$$  \hfill (113) 

and that $\{f_k\}_{k \in K}$ **converges uniformly** to $f$ and write $f_k \Rightarrow f$ if

$$\forall V \in V \ \exists k_0 \in K \ \forall x \in X : k \geq k_0 \implies (f_k(x), f(x)) \in V$$  \hfill (114)
In the special case where $X$ is a topological space with topology $\mathcal{T}$, we call the sequence $\{f_k\}_{k \in \mathbb{X}}$ \textbf{locally uniformly convergent} (see [Pro17b]) if every point in $S$ has a neighborhood in which the sequence converges uniformly. Symbolically,

$$\forall V \in \forall x_0 \in S \exists A \in T(x_0) \exists k_0 \in \mathbb{X} \forall x \in A : k \geq k_0 \implies (f_k(x), f(x)) \in V. \quad (115)$$

If the index $k_0$ does not depend on the neighborhood $A$ and the point $x_0$, then this is equivalent to uniform convergence. It is still more powerful than pointwise convergence. For example, power series are locally uniformly convergent on the interior of their domain of convergence - see proposition 132.

A slightly weaker notion is that of \textbf{compact convergence} (see [Pro17a]), which is defined as uniform convergence on any compact subset. Symbolically,

$$\forall V \in \forall \text{ compact } C \subseteq S \exists k_0 \in \mathbb{X} \forall x \in C : k \geq k_0 \implies (f_k(x), f(x)) \in V. \quad (116)$$

\textbf{Definition 345.} Fix two uniform spaces $(X, \mathcal{U})$ and $(Y, \mathcal{V})$. We say that the function $f : X \to Y$ \textbf{is uniformly continuous} on the set $S \subseteq X$ if

$$\forall V \in \forall U \in \forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V. \quad (117)$$

Compare this to \textbf{pointwise continuity} on $S$, which is defined by eq. (112) as convergence for any $x_1 \in X$:

$$\forall V \in \forall x_1, x_2 \in S \exists U \in \forall (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V. \quad (118)$$

\textbf{Definition 346.} Fix a topological space $(X, \mathcal{T})$ and a uniform space $(Y, \mathcal{V})$. We say that the family $\mathcal{F}$ of functions from $X$ to $Y$ is \textbf{functionwise continuous} at $x_0 \in X$ if

$$\forall V \in \forall f \in F \exists A \in T(x_0) : f(A) \subseteq B(f(x_0), V), \quad (119)$$

and \textbf{equicontinuous} at $x_0 \in X$ if

$$\forall V \in \forall f \in F \exists A \in T(x_0) \forall f \in F : f(A) \subseteq B(f(x_0), V). \quad (120)$$

In the special case where $(X, \mathcal{U})$ is a uniform space, then we can define \textbf{uniform equicontinuity} of the family $\mathcal{F}$ on the set $S \subseteq X$ as

$$\forall V \in \forall U \in \forall f \in F \forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V \quad (121)$$

Compare this to \textbf{pointwise equicontinuity} of $\mathcal{F}$ on $S$, as defined by eq. (120) for all $x_1, x_2 \in S$,

$$\forall V \in \forall x_1, x_2 \in S \exists U \in \forall f \in F : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V \quad (122)$$
to functionwise uniform continuity of $\mathcal{F}$ on $S$, which is defined by eq. (117) for all $f \in F$,
\[ \forall V \in \mathcal{V} \ \forall f \in F \ \exists U \in \mathcal{U} : \forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V \quad (123) \]
and to functionwise pointwise continuity of $\mathcal{F}$ on $S$, i.e. regular continuity as defined by corollary 341 for all $x_1, x_2 \in S$ and all $f \in F$,
\[ \forall V \in \mathcal{V} \ \forall f \in F \ \forall (x_1, x_2) \in S : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in V \quad (124) \]

**Proposition 347.**

(a) A locally uniform limit of functions continuous at a point is continuous at that point.

(b) A uniform limit of functions uniformly continuous on a set is uniformly continuous on the set.

**Proof.** The two proofs are similar, but have a lot of subtle differences.

Fix uniform spaces $(X, \mathcal{U})$ and $(Y, \mathcal{V})$. Let $\{f_k\}_{k \in \mathcal{K}}$ be a net of functions from $S \subseteq X$ to $(Y, \mathcal{V})$.

**Proof of 347 (a).** Assume that the functions $f_k, k \in \mathcal{K}$ are continuous and that they converge to the function $f$ locally uniformly.

Fix an entourage $W \in \mathcal{V}$ and use definition 333 to obtain $V \subseteq W$ such that $3V \subseteq W$.

, and a point $x_0 \in S$. Let $A$ be a neighborhood of $x_0$. From locally uniform convergence, there exists an index $k_0 \in \mathcal{K}$ such that
\[ \forall k > k_0 \ \forall x \in A : (f_k(x), f(x)) \in V. \]

Fix $k > k_0$. From uniform continuity, there exists an entourage $U \in \mathcal{U}$ such that
\[ \forall x \in S : (x, x_0) \in U \implies (f_k(x_0), f_k(x)) \in V. \]

Combining the last two inequalities, we note that for any $x \in A$,

- $(f(x_0), f(x)) \in V$,
- $(f_k(x_0), f(x_0)) \in V$,
- $(f_k(x), f(x)) \in V$,

thus by applying the triangle inequality in proposition 332 twice, we obtain
\[ (f(x_0), f(x)) \in 3V \subseteq W \ \forall x \in A \cap B(x_0, U). \]

Given an entourage $W \in \mathcal{V}$, we found a neighborhood $A \cap B(x_0, U)$ of $x_0$ such that eq. (111) is satisfied. Thus, $f$ is continuous at $x_0$.  

---

165
Proof of 347 (a). Assume that the functions \(f_k, k \in \mathcal{K}\) are uniformly continuous and that they converge to \(f\) uniformly.

As in proposition 347 (a), fix entourages \(V, W \in V\) such that \(3V \subseteq W\). From uniform continuity,

\[
\forall k \in \mathcal{K} \exists U \in U \forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f_k(x_1), f_k(x_2)) \in V.
\]

From uniform convergence, there exists an index \(k_0 \in \mathcal{K}\) such that

\[
\forall k > k_0 \forall x \in S : (f_k(x), f(x)) \in V.
\]

Fix an index \(k > k_0\) and let \(U \in U\) be such that

\[
\forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f_k(x_1), f_k(x_2)) \in V. \tag{125}
\]

For any two points \(x_1, x_2 \in S\), we also have that

\[
(f(x_i), f_k(x_i)) \in V, i = 1, 2. \tag{126}
\]

Analogously to proposition 347 (a), from (125) and (126), we obtain

\[
\forall x_1, x_2 \in S : (x_1, x_2) \in U \implies (f(x_1), f(x_2)) \in 3V \subseteq W.
\]

Thus, the entourage \(U\) depends on \(W\) and not on \(x_1\) and \(x_2\). Technically, it also depends on \(k_0\), however we are only concerned with existence and not uniqueness. Hence, \(f\) is uniformly continuous.

\[
\square
\]

Definition 348. Uniform spaces and uniformly continuous functions form a subcategory of \(\text{Top}\) (see definition 296). We denote this category by \(\text{Met}\).

Definition 349. A net \(\{x_k\}_{k \in \mathcal{K}}\) in a uniform space \((X, \mathcal{V})\) is called a fundamental net or Cauchy net if

\[
\forall V \in \mathcal{V} \exists k_0 \in \mathcal{K} \forall k, m \geq k_0 : (x_k, x_m) \in V.
\]

Lemma 350. A net in a uniform space that has a limit point is fundamental.

Proof. If \(x_0\) is a limit point of the net \(\{x_k\}_{k \in \mathcal{K}}\), the net is eventually in every ball \(B(x_0, V)\), which implies definition 349.

\[
\square
\]

Definition 351. A uniform space is called complete if it is Hausdorff and if every fundamental net converges.

The completion of uniform space \((X, \mathcal{V})\) is a (uniformly continuous) embedding \(f : X \to Y\) into a complete uniform space \((Y, \mathcal{U})\) such that \(\text{img}(X)\) is dense in \(Y\).

Theorem 352 (Uniform space completion). Every uniform space has a unique (up to an isomorphism) completion.

See also theorem 372 (Metric space completion).
**Theorem 353** (Cauchy’s net convergence criterion). A net in a complete uniform space is convergent if and only if it fundamental.

Explicitly, a net \( \{x_k\}_{k \in \mathbb{K}} \) in a complete uniform space \((X, \mathcal{V})\) is convergent if and only if for every entourage \( V \in \mathcal{V} \) there exists an index \( k_0 \) such that

\[
(x_k, x_m) \in V \quad \forall k, m \geq k_0.
\]

**Proof.**

**Proof of sufficiency.** Given by lemma 350

**Proof of necessity.** Given by definition 351

\[\square\]
7. Metric spaces

Metric spaces are uniform topological spaces with a lot of desirable properties. Most of real analysis that does not rely on certain algebraic structures generalizes well to metric spaces.
7.1. Metric topology

Definition 354. A **metric space** is a set $X$ along with a nonnegative real-valued function $ho : X \times X \to [0, \infty)$, called a **metric**, also called the **distance function**, such that

1. $M1$ $\rho(x, y) = 0 \iff x = y$
2. $M2$ (symmetry) $\rho(x, y) = \rho(y, x)$
3. $M3$ (triangle inequality) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$

If instead of $M1$, we have the weaker condition

$M1' \ \forall x \in X, \rho(x, x) = 0$,

we call $\rho$ a **pseudometric** and $(X, \rho)$ a **pseudometric space**.

(a) If $A \subseteq X$ is a set, then the restriction $(A, \rho|_A)$ is a metric space and it is called a **subspace** of $X$.

(b) Define the function

\[
B : X \times (0, \infty) \to \text{pow}(X),
\]

\[
B(x, r) := \{y \in X : \rho(x, y) < r\}.
\]

The set $B(x, r)$ is called an **open ball** or simply a **ball** with center $x$ and radius $r$.

The ball $B = B(0, 1)$ is called the **unit ball**.

(c) The set

\[
B(x, r) := \text{cl}(B(x, r))
\]

is called the **closed ball** with center $x$ and radius $r$.

(d) The set

\[
S(x, r) := \text{fr} B(x, r)
\]

is called the **sphere** with center $x$ and radius $r$.

(e) A set $A \subseteq X$ is called **bounded** if it is contained in some ball $B(x, r)$.

(f) A sequence $\{x_k\}_{k=1}^{\infty} \subseteq X$ is called **bounded** if the corresponding set $\{x_k : k = 1, 2, \ldots\}$ is bounded.

(g) If every set is bounded, we say that the metric itself is bounded.

(h) We say that a function $f : S \to X$ from a set $S$ to a metric space $(X, \rho)$ is **bounded** if its image $f(S)$ is a bounded set in $(X, \rho)$.
(i) Define the function
\[ \text{diam} : \text{pow}(X) \to [0, \infty], \]
\[ \text{diam}(A) := \sup\{\rho(x, y) : x, y \in A\}, \]
whose values include the nonnegative extended real numbers.
If it exists, we call the number \( \text{diam}(A) \) the **diameter** of \( A \).

(j) Define the function
\[ \text{dist} : X \times \text{pow}(X) \to [0, \infty), \]
\[ \text{dist}(x, A) := \inf\{\rho(x, a) : a \in A\}. \]
We call the number \( \text{dist}(x, A) \) the **distance from the point** \( x \) **to the set** \( A \). We use the convention that the infimum of an empty set of real numbers is \( +\infty \), hence \( \text{dist}(x, \emptyset) = \infty \).

**Proposition 355.** Let \( (X, \rho) \) be a pseudometric space. Define the equivalence relation
\[ x \cong y \iff \rho(x, y) = 0. \]
Then the following metric on the quotient set \( M := X/\cong \)
\[ \rho : M \times M \to [0, \infty) \]
\[ \rho([x], [y]) := \rho(x, y) \]
is well-defined.

**Proof.** The function \( \rho \) is well-defined since, if \( x \) and \( y \) both belong to the same equivalence class \([x] \), then \( \rho(x) = \rho(y) \). Thus, \( \rho \) does not depend on the choice of representatives.
Additionally, \( \rho \) is a metric since \( \rho([x], [y]) = 0 \) implies that \([x] = [y] \), that is, \( \rho(x, y) = 0 \). \( \Box \)

**Proposition 356.** A set \( A \) in a metric space \( (X, \rho) \) is bounded if and only if the set \( \{\rho(a, b) : a, b \in A\} \) is bounded as a partially ordered set.

**Definition 357.** Let \( (X, \rho) \) be a metric space. We define the **metric topology** \( \mathcal{T} \), also called the **induced topology**, as the topology generated by the neighborhood system
\[ \mathcal{B}(x) := \{B(x, \frac{1}{n}) : n = 1, 2, 3, \ldots\}. \tag{127} \]
If for some topological space \( (X, \mathcal{T}) \) there exists a metric such that \( \mathcal{T} \) is its induced topology, we say that the topology \( \mathcal{T} \) is **metrizable**.
It is often conventional to consider the alternative (larger) base
\[ \mathcal{B}'(x) := \{B(x, \varepsilon) : \varepsilon > 0\}. \tag{128} \]
**Figure 5:** There is a nested ball around every point in an open ball

**Proof.** This is indeed a neighborhood system as it satisfies BP1-BP3:

**BP1** Every point $x$ belongs to any ball centered at $x$.

**BP2** Fix $x \in X$ and two balls $B(x, \frac{1}{n})$ and $B(x, \frac{1}{m})$. Then

$$B(x, \frac{1}{\max\{n,m\}}) \subseteq B(x, n) \cap B(x, m).$$

**BP3** Fix $x, y \in X$ and let $x \in B(y, \frac{1}{n})$, i.e. $\rho(x, y) < \frac{1}{n}$.

Define $m$ to be the smallest positive integer such that

$$\frac{1}{m} \leq \min\{\rho(x, \frac{1}{n}), \frac{1}{n} - \rho(x, \frac{1}{n})\}.$$  

Note that $m$ exists since the positive integers are well-founded.

Let $z \in B(x, \frac{1}{m})$. There are two cases:

- If $\rho(x, y) \leq \frac{1}{2n}$, then
  $$\rho(z, y) \leq \rho(z, x) + \rho(x, y) < \frac{1}{m} + \rho(x, y) \leq \frac{1}{n} + \frac{1}{n} \leq 2 \frac{1}{2n} = \frac{1}{n}.$$  

- If $\rho(x, y) > \frac{1}{2n}$, then
  $$\rho(z, y) \leq \rho(z, x) + \rho(x, y) < \frac{1}{m} + \rho(x, y) \leq (\frac{1}{n} - \rho(x, y)) + \rho(x, y) = \frac{1}{n}.$$  

In both cases, $B(x, \frac{1}{m}) \subseteq B(y, \frac{1}{n})$. 

\[\square\]

**Proposition 358.** The metric topology $\mathcal{T}$ on $X$ induced by $\rho$ has the following properties:

(a) For every point $x \in X$ and any radius $r > 0$, the ball $B(x, r)$ is an open set and, hence, a neighborhood of $x$.

(b) $\mathcal{T}$ is first-countable.

(c) $\mathcal{T}$ is sequential.

(d) $\mathcal{T}$ is Hausdorff.

**Proof.**
Proof of 358 (a). Obvious from definition 357.

Proof of 358 (b). Since definition 357 involves generating a topology using a neighborhood system of countable local neighborhoods, \( \mathcal{F} \) is first-countable.

Proof of 358 (c). Follows from proposition 358 (b) and proposition 287.

Proof of 358 (d). Let \( x, y \in X \) be distinct points. Define
\[
r := \frac{1}{2} \rho(x, y),
\]
so that
\[
B(x, r) \cap B(y, r) = \emptyset.
\]

\[ \square \]

Definition 359. Let \((X, \rho)\) be a metric space.

We define the metric uniformity \( V \), also called the induced uniformity, as the uniformity generated by the countable base
\[
\mathcal{B} := \{ V_n : n = 1, 2, \ldots \},
\]
where
\[
V_n := \rho^{-1}([0, \frac{1}{n})).
\]

As for the metric topology, we can instead consider the base
\[
\mathcal{B}' := \{ \rho^{-1}([0, \varepsilon)) : \varepsilon > 0 \}.
\]

Proof. Each relation \( V_n \) is obviously an entourage by \( M1 \) and \( M2 \). We will prove that \( \mathcal{B} \) is indeed a uniform space base.

BU1 For nonnegative integers \( n, m \) we have
\[
V_n \cap V_m = \{ (x, y) \in X \times X : \rho(x, y) < \frac{1}{n} \text{ and } \rho(x, y) < \frac{1}{m} \} = V_{\max(n, m)}.
\]

Pick any integer \( k \geq \max\{n, m\} \), so that
\[
V_k \subseteq V_n \cap V_m.
\]

BU2 Fix \( V_n \in V \) and \( m := 2n \). By the triangle inequality, we have that if \( \rho(x, y) < m \) and \( \rho(y, z) < m \), then
\[
\rho(x, z) \leq \rho(x, y) + \rho(y, z) < \frac{1}{m} + \frac{1}{m} = \frac{1}{n}.
\]

Thus,
\[
V_m + V_m = \{ (x, z) : \exists y \in X : \rho(x, y) < \frac{1}{m} \text{ and } \rho(y, z) < \frac{1}{m} \} \subseteq V_n
\]
Proposition 360. The metric topology and the uniform topology from the metric uniformity coincide.

Theorem 361. A uniform space $X$ is metrizable if and only if $w(X) \leq \aleph_0$. 

[Eng89] thm. 8.1.21

Definition 362. Let $(X, \rho)$ and $(Y, \nu)$ be two metric spaces. We say that the function $f : X \to Y$ is a distance preserving map or isometry or isometric embedding if

$$\forall x, y \in X, \rho(x, y) = \nu(f(x), f(y)).$$

If $f$ is bijective, we say that $X$ and $Y$ are isometric.

Proposition 363. An isometry $f : (X, \rho) \to (Y, \nu)$ is always injective.

Proof. If $f(x) = f(x')$, then by definition 354, $x = x'$.

Definition 364. Metric spaces and monotone maps form a subcategory of $\text{Unif}$ (see definition 348). We denote this category by $\text{Met}$.

Definition 365. Two metrics $\rho$ and $\nu$ on the set $X$ are said to be equivalent if $\rho$ and $\nu$ have the same metric topology. They are said to be strongly equivalent if there exist constants $\alpha, \beta \in \mathbb{R}$ such that for every $x, y \in X$ we have

$$\alpha \nu(x, y) \leq \rho(x, y) \leq \beta \nu(x, y).$$

Remark 366. All types of convergence from section 6.2 (Topological nets), section 6.4 (Topological continuity) and section 6.10 (Uniform spaces) hold in metric spaces using the metric topology and metric uniformity structure.

It is conventional to prefer the bases equ. (128) and equ. (130) to the bases equ. (127) and equ. (127).

For example, given two metric spaces $X$ and $Y$, continuity of $f : X \to Y$ at $x_0 \in X$ (see definition 290) is usually written using the “epsilon-delta notation” as

$$\forall \varepsilon > 0 \exists \delta > 0 : \rho_X(x, x_0) < \delta \implies \rho_Y(f(x), f(x_0)) < \varepsilon$$

for any $x \in X$.

Definition 367. A metric $\rho$ on a magma $G$ is said to be left translation-invariant if

$$\rho(ax, ay) = \rho(x, y) \ \forall a, x, y \in G$$

and right translation-invariant if

$$\rho(xa, ya) = \rho(x, y) \ \forall a, x, y \in G.$$
7.2. Complete metric spaces

**Definition 368.** A metric space is said to be **complete** if

(a) Every fundamental sequence converges.

(b) It is complete as a uniform space in the sense of definition 351

**Proof.** The equivalence is due to proposition 358 (c) and proposition 358 (d). \qed

**Proposition 369.** In a metric space, any **fundamental sequence** \( \{x_k\}_{k=1}^n \) is bounded.

**Proof.** Since the set

\[ I := \{x_k : k \leq k_0 \} \]

is finite, it has a finite diameter.

Fix \( \varepsilon > 0 \). Since the sequence is fundamental, there exists an index \( k_0 \) such that

\[ \rho(x_k, x_m) < \varepsilon \quad \forall k, m \geq k_0. \]

We are only interested in the case \( \rho(x_{k_0}, x_m) < \varepsilon \).

Let \( k < k_0 \) and \( m \geq k_0 \). Then

\[ \rho(x_k, x_m) \leq \rho(x_k, x_{k_0}) + \rho(x_{k_0}, x_m) < \text{diam}(I) + \varepsilon, \]

which is a finite number.

Thus, the distance between any two elements of the sequence is finite and the sequence is bounded. \qed

**Proposition 370.** In any metric space, a **fundamental sequence** converges to a value if and only if it has a subsequence that converges to the same value.

**Proof.** Let \( (X, \rho) \) be a metric space and let \( \{x_k\}_{k=1}^\infty \) be a fundamental sequence.

**Proof of necessity.** Obvious

**Proof of sufficiency.** Assume that the subsequence \( \{x_{k_n}\}_{n=1}^\infty \) converges to \( x \). Fix \( \varepsilon > 0 \). There exist \( k_0 \) and \( n_0 \) such that

\[ \rho(x_k, x_m) < \frac{\varepsilon}{2} \quad \forall k, m \geq k_0 \]

\[ \rho(x, x_{k_n}) < \frac{\varepsilon}{2} \quad \forall n \geq n_0. \]

Fix \( k \geq k_0 \) and let \( n \geq n_0 \) be such that \( k_n \geq k_0 \). Then

\[ \rho(x, x_k) \leq \rho(x, x_{k_n}) + \rho(x_{k_n}, x_k) < \varepsilon. \]

Since \( \varepsilon \) was arbitrary, we conclude that \( \lim_{k \to \infty} x_k = \lim_{n \to \infty} x_{k_n} = x. \) \qed

**Lemma 371.** Let \( X \) be a metric space. If both \( f : X \to Y \) and \( g : X \to Z \) are completions of \( X \), then \( Y \) and \( Z \) are isometric.
**Proof.** Let \( y \in Y \) and let \( \{x_k\}_{k=1}^{\infty} \subseteq X \) be a sequence such that 
\[
f(x_k) \xrightarrow[k \to \infty]{} y.
\]

Such a sequence exists since \( f(X) \) is dense in \( Y \).

Define \( z := \lim_{k \to \infty} g(x_k) \). Since both \( f \) and \( g \) are isometries, \( z \) does not depend on the choice of sequence \( \{x_k\}_{k=1}^{\infty} \) such that \( f(x_k) \to y \). Furthermore, if \( z \in Z \) is given rather than \( y \in Y \), an analogous process allows us to determine \( y \) uniquely based on \( z \).

Thus, we have a bijective isometry between \( Y \) and \( Z \).

**Theorem 372** (Metric space completion). Every metric space has a unique (up to an isometry) completion. This is a special case of theorem 352 (Uniform space completion) that we prove fully.

**Proof.** Let \((X, \rho)\) be a metric space. Uniqueness of the completion follows from lemma 371. We will only show existence.

(a) First, we build the pseudometric space \((F, \rho)\). We deal with fundamental sequences and isometries in pseudometric spaces, where the definitions, however, does not change.

Define \( F \) to be the set of all fundamental sequences in \( X \). Define the pseudometric
\[
\rho : F \times F \to \mathbb{R}_{\geq 0} \\
\rho \left( \{x_k\}_{k=1}^{\infty}, \{y_k\}_{k=1}^{\infty} \right) := \lim_{k \to \infty} \rho(x_k, y_k).
\]

We first show that is well-defined as a function. Let \( \{x_k\}_{k=1}^{\infty} \) and \( \{y_k\}_{k=1}^{\infty} \) be two sequences. Fix \( \varepsilon > 0 \). Then there exists an \( k_0 \) such that
\[
\rho(x_k, x_k) < \frac{\varepsilon}{2} \text{ and } \rho(y_k, y_m) < \forall k, m \geq k_0 \frac{\varepsilon}{2}.
\]

Fix \( k, m \geq k_0 \). Then
\[
\rho(x_k, y_k) \leq \rho(x_k, x_m) + \rho(x_m, y_m) + \rho(y_m, y_k) < \rho(x_m, y_m) + \varepsilon,
\]
hence
\[
|\rho(x_k, y_k) - \rho(x_m, y_m)| < \varepsilon.
\]

Thus, the sequence \( \{\rho(x_k, y_k)\}_{k=1}^{\infty} \) is fundamental and, by theorem 53, it is convergent. Now we check that \( \rho \) is indeed a pseudometric:

(b) **Proof of M1’.** For every sequence \( x \in F \),
\[
\rho(x, x) = \lim_{k \to \infty} \rho(x_k, x_k) = 0.
\]

(c) **Proof of M2.** For all sequences \( x, y \in F \),
\[
\rho(x, y) = \lim_{k \to \infty} \rho(x_k, y_k) = \lim_{k \to \infty} \rho(y_k, x_k) = \rho(y, x).
\]

175
(d) **Proof of M3.** For all sequences \( x, y, z \in F \),

\[
\rho(x, z) = \lim_{k \to \infty} \rho(x_k, z) \leq \lim_{k \to \infty} \rho(x_k, y_k) + \lim_{k \to \infty} \rho(y_k, z) = \rho(x, y) + \rho(y, z).
\]

(e) We prove that every fundamental sequence in \((F, \rho)\) is convergent.

Let \( \{c^{(k)}\}_{k=1}^{\infty} \) be a fundamental sequence (of sequences) in \((F, \rho)\). Thus, for every \( k = 1, 2, ... \), there exists an index \( n_k \) such that

\[
\rho(c^{(k)}_m, c^{(k)}_{n_k}) < \frac{1}{k} \quad \forall \ m \geq n_k.
\]

Define the sequence

\[
d_k := c^{(k)}_{n_k}, \quad k = 1, 2, ...
\]

To see that it is fundamental, fix \( \varepsilon > 0 \). Now since the sequence \( \{c^{(k)}\} \) in \( F \) is fundamental, there exists \( k_0 \) such that

\[
\rho(c^{(k)}_m, c^{(m)}_l) = \lim_{k \to \infty} \rho(c^{(k)}_k, c^{(m)}_l) < \frac{\varepsilon}{2} \quad \forall k, m \geq k_0.
\]

Let \( m_0 \geq k_0 \) be an index such that

\[
\frac{2}{m_0} < \frac{\varepsilon}{2}.
\]

Fix \( k \geq m \geq m_0 \). Let \( l \geq \max\{n_k, n_m\} \) be such that

\[
\rho(c^{(k)}_l, c^{(m)}_l) < \frac{\varepsilon}{2}.
\]

Then

\[
\rho(d_k, d_m) = \rho(c^{(k)}_{n_k}, c^{(m)}_{n_m}) \leq \rho(c^{(k)}_{n_k}, c^{(k)}_l) + \rho(c^{(k)}_l, c^{(m)}_l) + \rho(c^{(m)}_l, c^{(m)}_{n_m}) \leq \frac{1}{k} + \frac{\varepsilon}{2} + \frac{1}{m} \leq \frac{2}{m_0} + \frac{\varepsilon}{2} < \varepsilon.
\]

Thus, we have

\[
\rho(d_k, d_m) < \varepsilon \quad \forall k \geq m \geq m_0,
\]

which proves that the sequence \( \{d_k\}_{k=1}^{\infty} \) is fundamental in \((X, \rho)\).

Now it remains to show that \( c^{(k)}_{k=1} \) is fundamental in \((F, \rho)\).

Fix \( \varepsilon > 0 \) and let \( k_0 \) be such that

\[
\frac{1}{k_0} \leq \frac{\varepsilon}{2}.
\]
and
\[ \rho(d_k, d_m) < \frac{\varepsilon}{2} \quad \forall k, m \geq k_0. \]

Now fix \( i \geq k_0 \). We have, for all \( k \geq i \),
\[
\rho(c_n^{(k)}, d_k) = \rho(c_n^{(k)}, c_n^{(k)}) \leq \\
\leq \rho(c_n^{(k)}, c_n^{(k)}) + \rho(c_n^{(k)}, d_k) = \\
= \rho(c_n^{(k)}, c_n^{(k)}) + \rho(d_k, d_k) < \\
< \frac{1}{k} + \frac{\varepsilon}{2} < \varepsilon.
\]

Hence,
\[
\rho(c^{(k)}, d) = \lim_{k \to \infty} \rho(c^{(k)}, d_k) = \lim_{k \to \infty} \rho(c^{(k)}, c_n^{(k)}) < \varepsilon.
\]

Thus, given \( \varepsilon > 0 \), we found an index \( k_0 \) such that
\[
\rho(c^{(k)}, d) < \varepsilon \quad \forall g \geq k_0.
\]

Thus, \( d = \lim_{k \to \infty} c^{(k)} \) and \((F, \rho)\) is a complete pseudometric space.

(f) We construct an isometry of \((X, \rho)\) into \((F, \rho)\).

Define the function
\[
t : X \to F \\
t(x) : = (x, x, x, ...),
\]
which sends each element of \( X \) into the corresponding constant sequence in \( F \).

It is an isometry since
\[
\rho(t(x), t(y)) = \lim_{k \to \infty} \rho(x, y) = \rho(x, y).
\]

(g) We show that the image \( t(X) \) is dense in \((F, \rho)\).

Fix the fundamental sequence \( y := \{y_k\}_{k=1}^{\infty} \). Define the sequence \( x \) of sequences
\[
x^{(k)} := t(y_k), l = 1, 2, ...
\]
It is fundamental in \((F, \rho)\) since \( e \) is an isometry and since \( y \) is fundamental in \((X, \rho)\).

Fix \( \varepsilon > 0 \). Let \( k_0 \) be such that
\[
\rho(y_k, y_m) < \varepsilon \quad \forall k, m \geq k_0.
\]
For $i, k \geq k_0$, we have
\[ \rho(x^{(k)}_k, y_k) \leq \rho(x^{(k)}_k, y_k) + \rho(y_k, y_k) = 0 + \rho(y_k, y_k) < \varepsilon, \]

hence
\[ \rho(x^{(k)}_k, y) = \lim_{k \to \infty} \rho(x^{(k)}_k, y_k) < \varepsilon. \]

We conclude that $x^{(k)}_k \longrightarrow y$ in $(F, \rho)$, which implies that $e(X)$ is dense in $(F, \rho)$.

(h) We build a complete metric space $(C, \nu)$ from $(F, \rho)$.

We use proposition 355 to construct a complete metric space $(C, \nu)$ from the complete pseudometric space $(F, \rho)$.

We adapt $\iota$ to the equivalence classes on $C$:

\[ \hat{\iota} : X \to C \]
\[ \hat{\iota}(x) := [\iota(x)]. \]

Thus, $\iota$ embeds $X$ into the complete metric space $C$. \hfill \square

**Theorem 373** (Cantor’s nested compact theorem). *A descending sequence of nonempty compact sets $F_1 \supseteq F_2 \supseteq \ldots$ in a complete metric space such that $\text{diam}(F_i) \to 0$ intersects at exactly one point (compare with theorem 386 (Kuratowski’s noncompactness lemma)).*

**Proof.** Choose an element $x_k \in F_k$ for any $i = 1, 2, \ldots$. Then the sequence $\{x_k\}_{k=1}^{\infty}$ is fundamental. To see this, let $\varepsilon > 0$ and let $k_0$ be an index such that $\text{diam}(F_{k_0}) < \varepsilon$. Then if $j \geq i \geq k_0$, $x_m$ is contained in $F_k$ and $\rho(x_k, x_m) < \varepsilon$. Thus, the sequence is indeed fundamental and, since the space is complete, it has a limit point $x$.

The point $x$ is contained in every set $F_k, i = 1, 2, \ldots$ since all of the sets $F_k$ are closed (by corollary 383) and contain their limit points. Thus,

\[ x \in \bigcap_{k=1}^{\infty} F_k. \]

Furthermore,

\[ \text{diam} \left( \bigcap_{k=1}^{\infty} F_k \right) = 0, \]

hence $x$ is the only point in the intersection. \hfill \square
7.3. Hausdorff distance

Let $(X, \mu)$ be a complete metric space.

**Definition 374.** Fix two sets $E \subseteq X$ and $F \subseteq X$.

The **excess** of $E$ beyond $F$ is defined as

$$
e : \text{pow} X \times \text{pow} X \to \mathbb{R} \cup \{\infty\}$$

$$e(E, F) :=
\begin{cases}
+\infty, & E = \emptyset, D = \emptyset \\
0, & E = \emptyset, D \neq \emptyset \\
\sup_{x \in E} \text{dist}(x, F) \overset{\delta}{=} \inf\{\delta > 0 : E \subseteq F_\delta\}, & E \neq \emptyset
\end{cases}$$

where $F_\delta := \{y \in X : \text{dist}(y, F) \leq \delta\}$.

The **Pompeiu-Hausdorff distance** or simply **Hausdorff** distance between them is then defined as

$$h(E, F) := \max\{e(E, F), e(F, E)\} = \inf\{\delta > 0 : E \subseteq F_\delta, F \subseteq E_\delta\}.$$ 

**Proof.** (of the equality *) Note that the set

$$F_{e(E, F)} = \{x \in X : \text{dist}(x, F) \leq \sup_{x \in E} \text{dist}(x, F)\}$$

obviously includes $E$.

Now let $\delta > 0$ be any real number that satisfies $E \subseteq F_\delta$, i.e.

$$E \subseteq F_\delta = \{x \in X : \text{dist}(x, F) \leq \delta\},$$

which implies that

$$e(E, F) = \sup_{x \in E} \text{dist}(x, F) \leq \delta.$$ 

\[ \square \]

**Proposition 375.** The Hausdorff distance is a metric on the nonempty compact subsets of $X$.

**Proof.** Let $E$, $F$ and $G$ be nonempty compact subsets of $X$.

The function $h$ is nonnegative. Since we exclude empty and unbounded sets, We do not care about infinite values.

**Proof of M1.** Obviously $h(E, E) = 0$. If $h(E, F) = 0$, then there exists no point of $E$ outside $F$ and vice versa, hence $E = F$.

**Proof of M2.** This follows from the symmetry of the max function.

**Proof of M3.** For any point $y \in X$, we have

$$\text{dist}(x, G) = \inf_{z \in G} \mu(x, z) \leq \mu(x, y) + \inf_{y \in G} \mu(y, z) = \mu(x, y) + \text{dist}(y, G).$$
Select $y \in F$ that minimizes the distance $\mu(x, y)$ over $F$ (compactness allows us), so that

**TODO:** Prove Weierstrass’ theorem

$$\text{dist}(x, G) \leq \mu(x, y) + \text{dist}(y, G) = \text{dist}(x, F) + \text{dist}(y, G) \leq \text{dist}(x, F) + e(F, G).$$

It now follows that

$$e(E, G) = \inf\{\delta > 0 : E \subseteq G_{\delta}\} =$$

$$= \inf\{\delta > 0 : E \subseteq \{x \in X : \text{dist}(x, G) \leq \delta\}\} \leq$$

$$\leq \inf\{\delta > 0 : E \subseteq \{x \in X : \text{dist}(x, F) + e(F, G) \leq \delta, y \in X\}\} =$$

$$= e(F, G) + \inf\{\delta > 0 : E \subseteq F_{\delta}\} =$$

$$= e(F, G) + e(E, F).$$

\[\square\]
7.4. Totally bounded sets

Let \((X, \rho)\) be a metric space.

**Definition 376.** We say that \(E \subseteq X\) is an \(\varepsilon\)-net for the set \(A \subseteq X\) if
\[
A \subseteq \bigcup_{x \in E} B(x, \varepsilon). \tag{132}
\]

**Definition 377.** The space \(A \subseteq X\) is called **totally bounded** if any of the following equivalent conditions hold:

(a) For every \(\varepsilon > 0\) there exists a finite cover of \(A\) with sets with diameter at most \(\varepsilon\).
(b) For every \(\varepsilon > 0\) there exists a finite \(\varepsilon\)-net of \(A\).
(c) Kuratowski’s noncompactness measure \(\alpha(A)\) is zero.
(d) The ball noncompactness measure \(\beta(A)\) is zero.
(e) Every sequence in \(A\) admits a fundamental subsequence.

Totally bounded sets are sometimes called **precompact** because of theorem 382. This equivalence requires the metric space to be complete, however.

**Proof.**

**Proof of equivalence of 377 (a) and 377 (c).** Straightforward.

**Proof of equivalence of 377 (b) and 377 (d).** Straightforward.

**Proof that 377 (b) implies 377 (a).** Given \(\varepsilon > 0\), any cover of \(A\) with balls of radius \(\frac{\varepsilon}{2}\) is a cover with sets of diameter \(\varepsilon\).

**Proof that 377 (a) implies 377 (b).** Fix \(\varepsilon > 0\) and \(\rho \in (0, \varepsilon)\) and let \(A_1, \ldots, A_n \subseteq \text{pow} X\) be a finite cover of \(A\) with sets of diameter at most \(\rho\).

Choose a point \(x_k\) from every \(A_k, k = 1, \ldots, n\). We then have that for every \(k = 1, \ldots, n\),
\[
A_k \subseteq \text{cl} B(x_k, \rho) \subseteq B(x_k, \varepsilon)
\]
\[
\implies A \subseteq \bigcup_{k=1}^n A_k \subseteq \bigcup_{k=1}^n B(x_k, \rho) \subseteq \bigcup_{k=1}^n B(x_k, \varepsilon),
\]
hence \(x_1, \ldots, x_n\) are centers of \(\varepsilon\)-balls that cover \(A\).

**Proof that 377 (b) implies 377 (e).** Let \(\{x_n\} \subseteq A\) be any sequence.

If we assume that \(\{x_n\}\) has no fundamental subsequence, then there exists \(\varepsilon_0 > 0\) such that \(\rho(x_k, x_m) > \varepsilon_0\) for any \(n, m \in \mathbb{Z}_{>0}\).

Consider a finite cover of \(A\) with \(\varepsilon_0\)-balls. By the pigeonhole principle, at least one of the balls contains more than one element of the sequence, which contradicts the assumption that all elements of the sequence have a distance of at least \(\varepsilon_0\).

Hence, an arbitrary sequence in \(A\) has a fundamental subsequence.
Proof that 377 (e) implies 377 (b). Assume that there exists $\varepsilon_0 > 0$, such that $A$ admits no finite cover by $\varepsilon_0$-balls.

Define $x_1 \in X, x_2 \in X \setminus B(x_1, \varepsilon_0), \ldots$ so that every two elements of the sequence $\{x_n\}$ have a distance of at least $\varepsilon_0$. But then the sequence is does not admit a fundamental subsequence, which contradicts our assumption.

This contradiction proves that $A$ admits a finite cover by $\varepsilon$-balls for every $\varepsilon > 0$. \hfill \Box

Corollary 378. Assume that $X$ is complete. The set $A \subseteq X$ is sequentially compact if and only if it is closed and totally bounded.

Proof. The property that every sequence has a fundamental subsequence is equivalent to sequential compactness for a closed set in a complete metric space. \hfill \Box

Proposition 379. Totally bounded sets are bounded.

Proof. Fix a totally bounded set $A \subseteq X$. Let $\varepsilon > 0$ and let $x_1, x_2, \ldots, x_n$ be a finite $\varepsilon$-net of $A$. The distance between two points of the $\varepsilon$-net is at most $2\varepsilon$. Then

$$A \subseteq \bigcup_{i=1}^{n} B(x_i, \varepsilon) \subseteq B(x_1, 2n\varepsilon).$$

Hence, $A$ is bounded. \hfill \Box

Proposition 380. If a set $A \subseteq X$ is totally bounded, then, so is its closure $\text{cl} \ A$.

Proof. Let $\varepsilon > 0$ and $\rho \in (0, \varepsilon)$ and let $x_1, \ldots, x_n \in X$ be the centers of a cover of $A$ with $\rho$-balls.

If $y$ is a point in $\text{cl} \ A \setminus A$, there exists a point $z \in A$ with $\rho(y, z) < \varepsilon - \rho$. Let $x_k \in A$ be one of the centers whose $\rho$-balls contain $z$. We then have that $y \in B(x_k, \varepsilon)$ since

$$\rho(x_k, z) \leq \rho(x_k, y) + \rho(y, z) < \rho + \varepsilon - \rho = \varepsilon.$$

Hence, the balls $\text{cl} \ B(x_k, \varepsilon)$ cover $\text{cl} \ A$, i.e.

$$\text{cl} \ A \subseteq \bigcup_{k=1}^{n} B(x_k, \varepsilon).$$

\hfill \Box

Lemma 381 (Lebesgue’s covering lemma). Assume that $X$ is complete. Let $A \subseteq X$ be sequentially compact. Given an open cover $\mathcal{F} \subseteq \text{pow} \ A$, there exists a number $\delta > 0$ such that every $\delta$-ball with a center in $A$ is contained in some set of the cover $\mathcal{F}$.

Proof. Assume that no such number $\delta > 0$ exists. Then for any natural number $n \in \mathbb{Z}_{>0}$, there exists an element $x_n \in A$ such that the ball $B(x_n, \frac{1}{n})$ is not contained in any set of the cover $\mathcal{F}$. Since $A$ is sequentially compact, the sequence $\{x_n\}_n$ contains a convergent subsequence $\{x_{n_k}\}_k$. 182
Define
\[ x := \lim_{k \to \infty} x_{n_k}. \]

Let \( E \) be a set in \( \mathcal{F} \) that contains \( x \). Since \( E \) is open, there exists some radius \( r > 0 \) such that \( B(x, r) \subseteq E \).

Choose any \( k_0 > \frac{2}{r} \) such that \( \rho(x_{n_{k_0}}, x) < \frac{r}{2} \). By the triangle inequality,
\[
B \left( x_{n_{k_0}}, \frac{r}{2} \right) \subseteq B(x, r) \subseteq E,
\]
which contradicts the choice of the sequence \( \{x_n\}_n \).

Hence, there exists a \( \delta > 0 \) such that for every \( x \in A \), the ball \( B(x, \delta) \) is contained in some element \( E \) of the cover \( \mathcal{F} \).

**Theorem 382.** Assume that \( X \) is complete. The set \( A \subseteq X \) is compact if and only if it is sequentially compact.

**Proof.**

**Proof of sufficiency.** Let \( \mathcal{F} \subseteq \text{pow } X \) be an open cover of \( A \).

By **Lemma 381** (Lebesgue’s covering lemma), there exists \( \delta > 0 \) such that for every \( x \in A \), the ball \( B(x, \delta) \) is contained in some set of the cover \( \mathcal{F} \). Let \( x_1, \ldots, x_n \) be a cover of \( A \) with \( \delta \)-balls.

For each \( k = 1, \ldots, n \) we have that the ball \( B(x_k, \delta) \) is contained in some set \( E_k \in \mathcal{F} \). Hence, \( E_1, \ldots, E_n \) is a finite subcover of \( A \), because
\[
A \subseteq \bigcup_{k=1}^{\infty} B(x_k, \delta) \subseteq \bigcup_{k=1}^{\infty} E_k.
\]

Thus, \( A \) is compact.

**Proof of necessity.** Let \( A \) be compact. Fix \( \varepsilon > 0 \) and take the cover
\[
\mathcal{F} := \{ B(a, \varepsilon) : a \in A \}.
\]

By compactness of \( A \), there exists a finite subcover. Thus, a finite cover of \( A \) with \( \varepsilon \)-balls exists for every \( \varepsilon > 0 \). **Definition 377** then implies that total boundedness is equivalent to sequential compactness because \( X \) is complete and \( A \) is closed.

**Corollary 383.** The following are equivalent for a set \( A \) in complete metric space:

(a) \( A \) is compact

(b) \( A \) is sequentially compact.

(c) \( A \) is closed and totally bounded.

**Proof.**

**Proof of equivalence of 383 (a) and 383 (b).** The equivalence is given by **Theorem 382**.

**Proof that 383 (b) implies 383 (c).** The equivalence is given by **Corollary 378**.
Figure 6: The operator $T_{\varepsilon}$ adds "spikes" to functions.

7.5. Noncompactness measures

**Definition 384.** Let $(X, \rho)$ be a metric space and let $\mathcal{B}$ be the family of bounded sets in $X$. We define the following functions:

(a) The **Kuratowski measure of noncompactness**, 
$$\alpha : \mathcal{B} \to [0, \infty)$$
$$\alpha(A) := \inf \{d > 0 : \exists U_1, \ldots, U_n \subseteq X : \text{diam } U_k < d \text{ and } A \subseteq \bigcup_{k=1}^{n} U_k \}$$

(b) The **ball measure of noncompactness**, 
$$\beta : \mathcal{B} \to [0, \infty)$$
$$\beta(A) := \inf \{r > 0 : \exists x_1, \ldots, x_2 \in X : A \subseteq \bigcup_{k=1}^{n} B(x_k, r) \}$$

**Example 385.** Consider the subsets $B_2 \subseteq B_3 \subseteq B_1 \subseteq C([0, 1])$, defined by

- $B_1 := \left\{ x \in C([0, 1]) : \begin{align*} 0 \leq t \leq 1 &\implies 0 \leq x(t) \leq 1 \\ x(0) = 0, x(1) = 1 \end{align*} \right\}$
- $B_2 := \left\{ x \in B_1 : \begin{align*} 0 \leq t \leq \frac{1}{2} &\implies 0 \leq x(t) \leq \frac{1}{2} \\ \frac{1}{2} \leq t \leq 1 &\implies \frac{1}{2} \leq x(t) \leq 1 \end{align*} \right\}$
- $B_3 := \left\{ x \in B_1 : \begin{align*} 0 \leq t \leq \frac{1}{2} &\implies 0 \leq x(t) \leq \frac{2}{3} \\ \frac{1}{2} \leq t \leq 1 &\implies \frac{1}{3} \leq x(t) \leq 1 \end{align*} \right\}$

Then $\alpha(B_1) = 1$, $\alpha(B_2) = \frac{1}{2}$, $\alpha(B_3) = \frac{1}{3}$ and $\beta(B_1) = \beta(B_2) = \beta(B_3) = \frac{1}{2}$.

**Proof.** Since the distance between any two functions from $B_1$ is at most 1, we have that $\text{diam } B_1 = 1$ and $\alpha(B_1) \leq 1$.

Fix $\varepsilon > 0$. For any function $f \in B_1$, continuity of $f$ gives us a radius $\delta_f > 0$ such that
$$x < 2\delta_f \implies f(x) < \varepsilon.$$  

Define
$$T_{\varepsilon}(f)(x) := \begin{cases} \frac{x}{\delta_f}, & 0 \leq x < \delta_f \\ f(\delta_f) + (1 - f(\delta_f))(2 - \frac{x}{\delta_f}), & \delta_f \leq x < 2\delta_f \\ f(x), & x \geq 2\delta_f, \end{cases}$$

184
so that
\[ ||T_\varepsilon(f) - f|| \geq T_\varepsilon(f)(\delta_f) - f(\delta_f) = 1 - f(\delta_f) > 1 - \varepsilon. \]

Additionally, because \( \delta_{T_\varepsilon(f)} < \delta_f \), we have that \( f(\delta_{T_\varepsilon(f)}) < \varepsilon \) and
\[ ||T_\varepsilon(T_\varepsilon(f)) - f|| \geq T_\varepsilon(T_\varepsilon(f))(\delta_{T_\varepsilon(f)}) - f(\delta_{T_\varepsilon(f)}) = 1 - f(\delta_{T_\varepsilon(f)}) > 1 - \varepsilon. \]

Thus, proceeding by induction, we see that for any \( m = 1, 2, ... \)
\[ ||T_\varepsilon^m(f) - f|| > 1 - \varepsilon, \]
where \( T_\varepsilon^m \) denotes repeated application of \( T_\varepsilon \).

Consider the sequence
\[ \{T_\varepsilon^k(f)\}_{k=0}^{\infty} = \{f, T_\varepsilon(f), T_\varepsilon(T_\varepsilon(f)), ...\}. \]

We can easily see that the distance between any two elements of the sequence, say \( T_\varepsilon^k(f) \) and \( T_\varepsilon^{k+m}(f) \), is strictly greater that \( 1 - \varepsilon \), i.e.
\[ ||T_\varepsilon^k(f) - T_\varepsilon^{k+m}(f)|| = ||T_\varepsilon^k(f) - T_\varepsilon^m(T_\varepsilon(f))|| > 1 - \varepsilon. \]

Hence, \( B_1 \) cannot be covered by a finite \((1 - \varepsilon)\)-net and \( \alpha(B_1) \geq 1 - \varepsilon \). Since \( \varepsilon > 0 \) can be made arbitrarily small, this implies that \( \alpha(B_1) \geq 1 \) and, because we already have the reverse inequality, \( \alpha(B_1) = 1 \).

In the set \( B_2 \), the maximum distance between two functions is \( \frac{1}{2} \), thus \( \text{diam}(B_2) = \frac{1}{2} \) and \( \alpha(B_2) \leq \frac{1}{2} \). We can then define an operator similar to \( T_\varepsilon \) that creates “spikes” of height \( \frac{1}{2} \) to prove the reverse inequality, obtaining
\[ \alpha(B_2) = \frac{1}{2}. \]

Finally, the set \( B_3 \) has diameter \( \frac{2}{3} \) and hence \( \alpha(B_3) = \frac{2}{3} \).

The ball measure for \( B_1 \) satisfies the inequalities
\[ \frac{1}{2} \leq \beta(B_1) \leq 1. \]

Additionally, \( B_1 \) is strictly contained in the ball centered in the constant function \( \frac{1}{2} \) with radius \( \frac{1}{2} \), which implies that \( \beta(B_1) \leq \frac{1}{2} \), hence \( \beta(B_1) = \frac{1}{2}. \)

For \( B_2 \) we have
\[ \frac{1}{4} \leq \beta(B_2) \leq \frac{1}{2}. \]

Assume that for some \( \varepsilon > 0 \) the set \( B_2 \) can be covered by a finite set of balls with centers \( \{f_1, ..., f_n\} \subseteq C([0, 1]) \) and radius \( \frac{1}{2} - \varepsilon \).
Figure 7: The function \( g \) always has points that are far enough from all \( f_k, k = 1, \ldots, n \).

Because of continuity, we can find a radius \( \delta > 0 \) such that for all \( f_k, k = 1, \ldots, n \) we have

\[
x \in \left[ \frac{1-\delta}{2}, \frac{1+\delta}{2} \right] \implies |f_k(x) - f_k(\frac{1}{2})| < \varepsilon.
\]

Consider the function

\[
g(x) := \begin{cases} 
0, & 0 \leq x < \frac{1-\delta}{2}, \\
\frac{2x+\delta-1}{2\delta}, & \frac{1-\delta}{2} \leq x \leq \frac{1+\delta}{2}, \\
1, & \frac{1+\delta}{2} < x \leq 1.
\end{cases}
\]

If \( f_k(\frac{1}{2}) \geq \frac{1}{2} \), then \( f_k(\frac{1-\delta}{2}) > \frac{1}{2} - \varepsilon \) and

\[
\|f_k - g\| \geq f_k(\frac{1-\delta}{2}) - g(\frac{1-\delta}{2}) = f_k(\frac{1-\delta}{2}) > \frac{1}{2} - \varepsilon.
\]

Analogously, if \( f_k(\frac{1}{2}) < \frac{1}{2} \), then \( f_k(\frac{1+\delta}{2}) < \frac{1}{2} + \varepsilon \) and

\[
\|g - f_k\| \geq g(\frac{1+\delta}{2}) - f_k(\frac{1+\delta}{2}) = 1 - f_k(\frac{1+\delta}{2}) > \frac{1}{2} - \varepsilon.
\]

Thus, for every \( k = 1, \ldots, n \) we have

\[
\|g - f_k\| > \frac{1}{2} - \varepsilon,
\]

i.e. \( g \) in not contained in a ball of radius \( \frac{1}{2} - \varepsilon \) around any of the centers \( f_1, \ldots, f_n \).

Hence, \( \beta(B_2) \geq \frac{1}{2} \), which implies \( \beta(B_1) = \frac{1}{2} \). Because of the inclusion \( B_2 \subsetneq B_3 \subsetneq B_1 \), we have

\[
\frac{1}{2} = \beta(B_2) \leq \beta(B_3) \leq \beta(B_1) = \frac{1}{2},
\]

hence \( \beta(B_1) = \frac{1}{2} \).

\[\square\]

**Theorem 386** (Kuratowski’s noncompactness lemma). *Let \( X \) be a Banach space and \( \{A_n\}_n \) be a decreasing sequence of nonempty closed subsets such that \( \alpha(A_n) \to 0 \). Then \( A := \bigcap_n A_n \) is nonempty and compact.*

**Proof.** The set \( A \) is compact because it is closed as the intersection of closed sets and \( \alpha(A) \leq \alpha(A_n) \to 0 \), hence \( \alpha(A) = 0 \).

It remains to show that \( A \) is nonempty. Choose any sequence \( \{x_n\}_n \) where \( x_n \in A_n \). Since any finite set is compact, we have that for any \( k \geq 1 \)

\[
\alpha(\{x_n_{n\geq k}\}) \leq \alpha(\{x_n_{n<k}\}) = \alpha(\{x_n\}_{n\geq k}) \leq \alpha(A_k) \to 0,
\]

hence the set \( \{x_n : n \geq 1\} \) is compact and thus sequentially compact. We can choose a convergent subsequence \( \{x_n_{k}\}_k \) of \( \{x_n\}_n \) whose limit lies in every \( A_n \) (since they are closed) and, consequently, in their intersection \( A \). So \( A \) is nonempty. \[\square\]
7.6. Lipschitz continuity

Definition 387. Let $f : X \to Y$ be a function between metric spaces.

(a) We say that $f : X \to Y$ is Hölder continuous at $x \in X$ with constant $L \geq 0$ and exponent $\alpha > 0$ if
$$\rho_Y(f(x_1), f(x_2)) \leq L \rho_X(x_1, x_2)^{\alpha} \quad \forall x_1, x_2 \in X.$$  
We refer to the smallest such constant, if any, as “the” Hölder constant.

(b) We say that $f$ is locally Hölder continuous if every point has a neighborhood where $f$ is Hölder continuous with the same exponent, but possibly with with a different constant.

(c) If $\alpha = 1$, we say that $f$ is Lipschitz continuous.

(d) If $X = Y$ and if $f$ is Lipschitz with constant $L < 1$, we call $f$ a contraction mapping.

(e) [DR14, p. 53] We say that $f$ is calm at $x$ if it satisfies the Lipschitz condition with one of the points fixed:
$$\rho_Y(f(x), f(x')) \leq L \rho_X(x, x') \quad \forall x' \in X.$$  

Proposition 388. A Hölder map is uniformly continuous.

Proof. Let $f : X \to Y$ be a Hölder map with constant $L$ and exponent $\alpha$.

Fix $\varepsilon > 0$. Then is enough to choose $\delta < \frac{\varepsilon}{\sqrt[\alpha]{L}}$, so that
$$\rho_X(x_1, x_2) < \delta \implies \rho_Y(f(x_1), f(x_2)) \leq L \rho_X(x_1, x_2)^{\alpha} < L \delta^\alpha < \varepsilon.$$

This implies uniform continuity. \qed

Corollary 389. A locally Hölder map is continuous.

Theorem 390 (Banach’s fixed point theorem). A contraction mapping in a complete metric space has a unique fixed point.

Proof. Let $f : X \to X$ be a contraction mapping. Fix any point $x_0 \in X$ and inductively define the sequence
$$x_{k+1} := f(x_k), \quad k = 1, 2, \ldots$$

Fix $\varepsilon > 0$. Since $L < 1$, there exists an index $k_0 > \log_L(\varepsilon)$ such that for positive integers $m$ and $k > k_0$,
$$\rho(x_k, x_{k+m}) = \rho(f^k(x_0), f^{k+m}(x_0)) \leq L^k \rho(x_0, x_m) < \varepsilon \rho(x_0, x_m).$$

187
Note that

\[ \rho(x_0, x_m) \leq \sum_{i=1}^{m} \rho(x_{i-1}, x_i) \leq \rho(x_0, x_1) \sum_{i=1}^{m} L^{i-1} = \rho(x_0, x_1) \frac{1 - L^m}{1 - L} \leq \rho(x_0, x_1) \frac{1}{1 - L}. \]

Thus,

\[ \rho(x_k, x_{k+m}) < \frac{\varepsilon \rho(x_0, x_1)}{1 - L}. \]

The constant on the right is linear in \( \varepsilon \) and does not depend on \( k \) or \( m \), hence \( \{x_k\}_{k=0}^{\infty} \) is a fundamental sequence. Since \( X \) is complete, the sequence has a limit \( x \).

Because of the continuity of \( f \) (see proposition 388),

\[ f(x) = f(\lim_{k \to \infty} x_k) = \lim_{k \to \infty} f(x_k) = \lim_{k \to \infty} x_k = x. \]
7.7. Function oscillation

**Definition 391.** Let \( X \) be a nonempty set and \((Y, \rho_Y)\) be a metric space. We define the oscillation of a function on a set as
\[
\omega : \text{fun}(X,Y) \times \text{pow}(X) \rightarrow [0, \infty)
\]
\[
\omega(f, A) := \sup \{ \rho_Y(f(x), f(y)) : (x, y) \in A \}.
\]

In particular, if \( X \) is itself a metric space, we define its modulus of continuity \( \omega(f, \delta) \) as the oscillation of \( f \) on the ball \( B(0, \delta) \).

**Proposition 392.** The modulus of continuity has the following basic properties:

(a) \( f \) is globally uniformly continuous if and only if for every \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that \( \omega(f, \delta) < \varepsilon \).

(b) \( \omega(f, \delta) \) is monotone in \( \delta \).

(c) For all \( \lambda, \delta > 0 \), we have the following analog of theorem 765 (Cauchy-Bunyakovsky-Schwarz inequality)
\[
\omega(f, \lambda \delta) \leq \omega(f, \lambda^2) + \omega(f, \delta^2).
\]

(d) For all \( \lambda, \delta > 0 \),
\[
\omega(f, \lambda \delta) \leq (\lambda + 1) \omega(f, \delta).
\]

**Proof.**

**Proof of 392 (a).** Follows directly from definition 345.

**Proof of 392 (b).** A supremum on a larger set is larger.

**Proof of 392 (c).** If \( \lambda \leq \delta \), clearly \( \lambda \delta \leq \delta^2 \). Otherwise, \( \lambda \delta < \lambda^2 \).

Combining the two inequalities with proposition 392 (b), we obtain eq. (133).

**Proof of 392 (d).** Note that
\[
\rho_X(x, y) < \delta \text{ implies } \rho_Y(f(x), f(y)) < \omega(f, \delta).
\]

We can multiply this by \( \lambda \) to obtain
\[
\lambda \rho_X(x, y) < \lambda \delta \text{ implies } \lambda \rho_Y(f(x), f(y)) < \lambda \omega(f, \delta).
\]

If \( \lambda \geq 1 \), then \( \rho_X(x, y) \leq \lambda \rho_X(x, y) \) and \( \rho_Y(f(x), f(y)) \leq \lambda \rho_Y(f(x), f(y)) \) and hence
\[
\omega(f, \lambda \delta) \leq \lambda \omega(f, \delta).
\]

Otherwise, \( \lambda < 1 \) and clearly \( \lambda \delta < \delta \), which by proposition 392 (b) implies
\[
\omega(f, \lambda \delta) \leq \omega(f, \delta).
\]

Combining the two cases, we obtain
\[
\omega(f, \lambda \delta) \leq \lambda \omega(f, \delta) + \omega(f, \delta),
\]

which we wanted to prove. \( \square \)
8. Geometry

Geometry is the multi-millennium evolution of attempts to measure parts of the earth. Ironically, it may be the main historical justification for the gradual axiomatization of mathematics. Completely abstract results about shapes date at least as early as in Ancient Greece. The important distinction between ancient geometry and modern geometry is the introduction of coordinates in the 17th century.

An axiomatic approach for a theory of plane and, solid figures was developed by Euclid in the third century BC. Later, Hilbert, Tarski and others independently proposed axioms that fit the requirements of modern logic systems. This is known today as synthetic Euclidean geometry and is mostly of theoretical interest because modern tools are easier to work with.

Descartes’ idea of coordinates connects problems of algebra and geometry in such a way that most of today’s mathematics seamlessly switches between algebraic and geometric interpretations of the same problem. The study of classical Greek geometry in terms of coordinates is known as analytic geometry.
8.1. Affine coordinate systems

Remark 393. Most humans possess a strong intuition for visual information like drawings or diagrams. A paper or a painting is only a medium for communicating information and emotions. Figure 8 contains some highlighted curves that our mind maps to abstract geometric figures, without considering the size limitations of the page, the precision of the drawings or the thickness of the lines.

Our goal is to map these visualizations to the concept of vector spaces. Formalisms at the level of formal logic will not be stated because we only want to sketch some high-level concepts. We only give definitions that are strictly necessary, plane geometry itself is described in section 8.3 (Analytic geometry in the plane). We will proceed as follows:

- Define an affine plane in definition 394 with auxiliary definitions.
- Describe the Euclidean plane $A_2$ in definition 395 as a very special affine plane.
- Give additional definitions for the Euclidean plane in definition 396.
- Define the set $F_2$ of free vectors over $A_2$ in definition 397.
- Show that $F_2$ is a two-dimensional vector space over $\mathbb{R}$ in theorem 398.
- Define coordinate systems that give explicit isomorphisms between $A_2$, $F_2$ and $\mathbb{R}^2$ in definition 399.
- Generalize these notions in remark 400

Definition 394. An affine plane consists of

- a set $X$, whose elements are called points,
- a family of subsets of $X$, whose members are called lines

with the additional relations

- a parallel relation $l \parallel g$ for lines that holds if either $l = g$ or if they have no points in common,
- a collinearity relation for a set $B$ of points that holds if $B$ is a subset of some line,

such that

A1 Given two distinct points, there exists only one line that contains both.

A2 Given a line $l$ and a point $P \notin l$, there exists exactly one line $g \parallel l$ that contains $P$.

TODO: Add diagram

Figure 8: A triangle, a circle and a line in the Euclidean plane.
Figure 9: Three lines and three points in the Euclidean plane. The lines $l$ and $g$ are collinear, while the point $R$ is between $P$ and $Q$.

Figure 10: Differently hatched half-planes in the Euclidean plane.

A3 There exist three non-collinear points.

**Definition 395.** The Euclidean plane $A_2$ is a formalization of a straight infinite surface. An axiomatic definition can be found in [nLa20c]. We will use that

- The Euclidean plane $A_2$ is an **affine plane**
- $A_2$ is a **complete metric space** with distance $\text{dist}$.
- There is a **betweenness** relation for points that says if the point $R$ is **between** $P$ and $Q$.

**Definition 396.** We will also need the following definitions:

(a) Every line $l$ gives rise to two (closed) **half-planes** $H^+$ and $H^-$ as follows:

- $H^+ \cap H^- = l$
- $H^+ \cup H^- = A_2$
- If $P \in H^+ \setminus l$ and $Q \in H^- \setminus l$, then there is a point $R \in l$ between $P$ and $Q$.

Note that the superscripts + and − are only for distinguishing between the two half-planes and are not assigned based on some property of the half-planes. See definition 413 for a definition of a half-plane that actually has a concept of signs.

(b) Every line $l$ and every point $R$ give rise to two (closed) **rays** $l^+$ and $l^-$ as follows:

- $l^+ \cap l^- = \{R\}$ are disjoint
- $l^+ \cup l^- = l$
- If $P \in l^+ \setminus \{R\}$ and $Q \in l^- \setminus \{R\}$, then $R$ is between $P$ and $Q$.

The rays $l^+$ and $l^-$ are called **opposite** of each other.

We say that $R$ is the **vertex** of $l^+$ and $l^-$. See definition 408 for a definition of a ray that actually has a concept of signs.
Figure 12: Bound vectors in the Euclidean plane can be regarded as oriented line segment.

(c) Two rays are said to be unidirectional if there exists a line distinct from the lines containing the rays, such that both rays are contained in the same half-plane with respect to the line.

(d) An ordered pair $\overrightarrow{PQ}$ of points is called a bound vector. The point $P$ is called the beginning of $\overrightarrow{PQ}$ and $Q$ is called the end of $\overrightarrow{PQ}$.

**Definition 397.** We say that the bound vectors $\overrightarrow{P_1Q_1}$ and $\overrightarrow{P_2Q_2}$ in $A_2$ are congruent if $\text{dist}(P_1, Q_1) = \text{dist}(P_2, Q_2)$ and if the rays $r_i, i = 1, 2$ beginning at $P_i$ and containing $Q_i$, are unidirectional.

We define free vectors as equivalence classes of bound vectors by this congruence relation. We denote the corresponding equivalence partition by $F_2$.

**Theorem 398.** The set $F_2$ of free vectors over $A_2$ is a two-dimensional vector space over $\mathbb{R}$ with the following operations:

(a) We define the sum of the cosets $[\overrightarrow{PQ}]$ and $[\overrightarrow{QR}]$ as the coset $[\overrightarrow{PR}]$.

(b) We define the scalar multiplication of $\lambda \in \mathbb{R}$ with the coset $[\overrightarrow{PQ}]$ to be the coset $[\overrightarrow{PR}]$, where $\overrightarrow{PR}$ is the unique vector that is unidirectional with $\overrightarrow{PQ}$ and $\text{dist}(P, R) = \lambda \text{dist}(P, Q)$.

**Proof.** Proving the well-definedness of the operations and verifying that $F_2$ is a two-dimensional vector space requires a lot of work and the proof is skipped.

**Definition 399.** Just because theorem 398 states that the set $F_2$ of free vectors is a vector space does not mean that we can work with it as with $\mathbb{R}^2$. ?? ([UNDEFINED]) says that $F_2$ is isomorphic to $\mathbb{R}^2$, however the proof requires the axiom of choice. The concrete way to select a basis in $F_2$ is through coordinate systems.

Somewhat confusingly, we define coordinate systems over $A_2$ rather than over $F_2$, but this will, soon be justified.

A coordinate system $Oxy$ in $A_2$ is a choice of

(a) A point $O \in A_2$, called the origin of the coordinate system.

(b) An ordered basis $(x, y)$ of $F_2$, called the basis of $Oxy$.

What we achieve through the choice of $O$ is that, for each point $P \in A_2$, we select the bound vector $\overrightarrow{OP} \in V_2$, called the radius vector of $P$. This injects $A_2$ into $V_2$, however if we take the free vector $[\overrightarrow{OP}]$, we instead obtain a bijection between $A_2$ and $F_2$.

Now that we have a correspondence between $A_2$ and $F_2$, coordinates for the point $P$ are defined simply as the coordinates of $[\overrightarrow{OP}]$ with respect to the basis $(x, y)$.

Thus, the pair $(A_2, Oxy)$ has an explicit isomorphism with $\mathbb{R}^2$.

The coordinate axis of $x$ is the unique ray starting at $O$ and containing the end of $x$. It is called the abscissa. The coordinate axis of $y$ is called the ordinate.
Remark 400. We sketched how to embed mental images of planes into $\mathbb{R}^2$, however in mathematics we are often interested in the opposite: given a set of points in $\mathbb{R}^2$, visualize them on a screen or paper and then absorb the resulting image in our brain.

This is one of the most powerful constructions in mathematics, yet it is, so intuitive that it is not really given a lot of attention, at least until generalizations are required. Given any vector space $V$ in the sense of definition 624, we want a way to assign a pair of numbers to each vector in $V$. This is only possible if $\dim V = 2$, however we can generalize this to tuples of coordinates via bases - see definition 614. This well for finitely dimensional vector spaces, however we need to generalize these notion for infinitely dimensional vector spaces and general modules over rings. This allows us to generalize coordinates further to manifolds - see definition 432.

See section 8.2 (Vector space geometry) for immediate generalizations of the concepts introduced here.
8.2. Vector space geometry

Remark 401. When speaking about vector spaces, we usually restrict ourselves to vector spaces over $\mathbb{R}$ or, at most, $\mathbb{C}$. This restriction may seem arbitrary, however important concepts like rays or convexity requires the field to be an extension of $\mathbb{R}$, and it just so happens that, by theorem 41 (Fundamental theorem of algebra) and proposition 708, the only nontrivial finite extension field of $\mathbb{R}$ is $\mathbb{C}$. It is technically possible to work with infinite extension fields, however in practice vector spaces over $\mathbb{C}$ are esoteric enough. A benefit of considering only $\mathbb{C}$ is given in remark 43.

Definition 402. A geometric shape is an informal notion that refers to certain special subsets of a vector space, usually defined in a coordinate-independent manner. Shapes in two-dimensional spaces are called figures and shapes in three dimensions are called surfaces.

When two geometric shapes intersect, we say that they are incident.

Definition 403. A point is a simple geometric shape comprising a singleton subset of any set (usually a vector space or a topological space). We use the convention remark 930 and, unless the distinction is important, we do not distinguish between singleton sets and their only element - e.g. in definition 418 (a).

Points are also called vectors, which is justified by definition 399.

Definition 404. The following bijections from $\mathbb{R}^n$ to $\mathbb{R}^n$ are collectively called Euclidean transformations or rigid motions in $\mathbb{R}^n$:

(a) For any vector $v \in \mathbb{R}^n$, the function

$$t_v(x) := x + v$$

is called a translation along the direction $v$.

This easily generalizes to an arbitrary magma $(M, \cdot)$ by setting

$$t_v(x) := v \cdot x$$

for some $v \in M$.

(b) For any scalar $\lambda \in \mathbb{R}^n$, the function

$$d_\lambda(x) := \lambda x$$

is called a dilation or scaling by $\lambda$.

This easily generalizes to an arbitrary (left) module $(M, +, \cdot)$.

(c) A composition of a dilation with a translation is called a homothety.

(d) For any point $p \in \mathbb{R}^n$ and any special (i.e. determinant one) orthogonal matrix $A \in O_n(\mathbb{R})$, the function

$$\text{rot}_{A,p}(x) := p + A(x - p)$$

is called a rotation by $A$ around $p$. 

195
(e) For any point \( p \in \mathbb{R}^n \), the function

\[
\text{ref}_p(x) := 2p - x
\]

is called a point reflection or inversion with respect to the point \( p \).

**Definition 405.** Let \( S \) be an arbitrary set and let \( M \) be a monoid with identity \( e \).

The **zero locus** or **set of zeros** of a function \( f : S \to M \) is the preimage

\[
f^{-1}(e) = \{ x \in X : f(x) = e \}.
\]

In practice, \( M \) is usually a ring or a module, in which case the zero locus is defined for the additive group, i.e.

\[
f^{-1}(0_M) = \{ x \in X : f(x) = 0_M \}.
\]

See also ?? ([UNDEFINED]), ?? ([UNDEFINED]) and ?? ([UNDEFINED]).

**Definition 406.** A **hypersurfaces** can have different meanings depending on the context.

We are interested in

(a) A parametric hypersurface (definition 294) is a purely topological definition.

(b) An affine variety (definition 436) is a purely algebraic definition.

(c) A manifold (definition 432) can be regarded as a geometric definition.

Note that all of the enumerated hypersurfaces have a concept of dimension. Hypersurfaces of dimension 2 are simply called **surfaces** (see definition 436 (e)) and hypersurfaces of dimension 1 are called **curves** (see also definition 436 (d) and definition 436 (d)).

**Definition 407.** A particularly important **curve** is a **line** in a vector space \( X \) over any field \( \mathbb{K} \), which can be defined equivalently as

(a) A subspace of \( X \) of dimension one. Note that this is not consistent with the other definitions because this defines only lines through the origin \( 0_X \). Hence, if \( L \subseteq X \) is a line (a subspace of dimension one) and if \( a \in X \) is any point, we define a line with origin \( a \) to be the translation \( a + L \).

(b) An algebraic curve in \( X \) given by a polynomial of degree one.

(c) If the field \( \mathbb{K} \) is ordered (usually when \( \mathbb{K} = \mathbb{R} \)), we can define a line with **directional vector** \( x \) and **origin** \( a \) as the parametric curve

\[
l : \mathbb{K} \to X \\
l(t) = tx + a.
\]
**Definition 408.** If \( X \) is a vector space over \( K \in \{\mathbb{R}, \mathbb{C}\} \), we define **closed rays** with a vertex \( a \) as the parametric curves

\[
\begin{align*}
    l^+ & : [0, \infty) \to X \\
    l^+(t) &= tx + a
\end{align*}
\]

and

\[
\begin{align*}
    l^- & : (\infty, 0] \to X \\
    l^-(t) &= tx + a
\end{align*}
\]

If the inequalities are strict, we instead obtain **open rays**.

Unless explicitly noted otherwise, we assume that the vertex of the ray is 0 because every ray is a translation of a ray centered at 0.

**Definition 409.** An open (resp. closed) **cone** in a vector space over \( K \in \{\mathbb{R}, \mathbb{C}\} \) is a union of open (resp. closed) rays with a common vertex, called the **vertex** of the cone.

**Definition 410.** Let \( M \) and \( N \) be two left \( R \)-modules. We say that the function \( f : M \to N \) is **affine** if it is a translation of a **linear function**, that is, if there exists a linear function \( l : M \to N \) and a constant \( a \in N \) such that \( f(x) = l(x) + a \).

**Definition 411.** Dually to **lines**, another particularly important **hypersurface** is a **hyperplane** in a vector space \( X \) over any field.

(a) A **linear hyperplane** is simply a subspace of \( X \) of codimension one. As in **definition 407 (a)**, we define a **affine hyperplane** to be a translation of a linear hyperplane.

(b) Linear hyperplanes (as defined in **definition 411 (a)**) are simply zero loci (kernels) of linear functionals. and affine hyperplanes are zero loci of affine functionals.

**Example 412.** Affine hyperplanes in \( \mathbb{R}^2 \) are **lines** and affine hyperplanes in \( \mathbb{R}^3 \) are planes.

Linear hyperplanes in \( \mathbb{R}^2 \) are the lines passing through the origin \((0,0)\) and linear hyperplanes in \( \mathbb{R}^3 \) are the planes incident to \((0,0,0)\).

**Definition 413.** Vector spaces over \( \mathbb{R} \) have the concept of **half-spaces**. Given a hyperplane \( H \) of the real vector space \( X \), defined by the affine functional \( l(x) = \langle x^*, x \rangle + a \), its closed half-spaces are defined as \( H^+ := \{l(x) \geq 0\} = \{\langle x^*, x \rangle \geq -a\} \) and \( H^- := \{l(x) \leq 0\} = \{\langle x^*, x \rangle \leq -a\} \).

If the inequalities are strict, we instead obtain **open half-spaces**.

**Definition 414.** A **polyhedron** in a real vector space is an intersection of half-spaces.
**Definition 415.** Although hyperplanes are defined for vector spaces over an arbitrary field \( \mathbb{K} \), we define hyperplane separation only for \( \mathbb{K} \in \{ \mathbb{R}, \mathbb{C} \} \) (see remark 401).

We say that the sets \( A, B \subseteq X \) are **separated** by the linear functional \( l \in X^* \) if there exists a real number \( c \in \mathbb{R} \) such that

\[
\text{real } l(x) < c \leq \text{real } l(y) \quad \forall x \in A, y \in B. \quad (135)
\]

See remark 43 for a justification of only considering the real part of \( l \).

The asymmetry in the inequalities eq. (135) can be inverted by considering \(-l(x)\) and \(-c\).

We say that \( A \) and \( B \) are **strongly separated** by \( l \) if both inequalities in eq. (135) are strict:

\[
\text{real } l(x) < c < \text{real } l(y) \quad \forall x \in A, y \in B. \quad (136)
\]

We can regard \( l \) as a hyperplane as in definition 411 (b), which justifies the terminology “hyperplane separation”. It is more correct, however, especially if \( \mathbb{K} = \mathbb{R} \), to say that they are separated by the affine hyperplane \( l(x) + c \).

If \( \mathbb{K} = \mathbb{R} \) eq. (135) is equivalent to requiring that \( A \) is contained in an open half-space relative to \( l(x) + c \) and that \( B \) is contained in the complementing closed half-space (or vice-versa). Equation (136) then states that both \( A \) and \( B \) are contained in opposite open half-spaces.

**Definition 416.**

(a) Given two points \( x, y \in X \) in a Banach space \( X \), we define the **line segment** between \( x \) and \( y \) as the parametric curve \( t \mapsto tx + (1-t)y, t \in [0,1] \). The image

\[
[x, y] := \{tx + (1-t)y : t \in [0,1]\}
\]

of this parametric curve is called the **convex hull** of \( x \) and \( y \). We usually use the term “line segment” to refer to the convex hull itself.

The length \( \text{len}([x, y]) \) of a line segment is defined as \( ||x - y|| \).

(b) We define the convex hull \( \text{conv } A \) of a set \( A \subseteq X \) as the union of all line segments with endpoints in \( A \).

(c) We call a set **convex** if it coincides with its convex hull, that is, if it contains the line segment between any two of its points.

**Proposition 417.** **Convex sets have the following basic properties:**

(a) A convex set is closed under convex combinations.

(b) A closed convex cone is closed under conic combinations.

(c) Any intersection of convex sets is convex.

*Proof.*
**Proof of 417 (a).** Fix a convex set \( C \). Let \( \sum_{k=1}^{n} t_k x_k \) be a convex combination of elements of \( C \).

We will use induction on \( n \). If \( n = 1 \), this is obvious. If \( n = 2 \), this is given by definition. Assume that it is true for \( n - 1 \). Denote \( s := \sum_{k=1}^{n-1} t_k \). If \( s = 0 \), take another convex combination in order to handle all the possible cases of the induction. Suppose \( s \neq 0 \). Then

\[
\sum_{k=1}^{n} t_k x_k = s \sum_{k=1}^{n} \frac{t_k}{s} x_k = s \sum_{k=1}^{n-1} \frac{t_k}{s} x_k + t_n x_n.
\]

By the inductive hypothesis, \( y \in C \). Note that \( s \in [0, 1] \) and that \( s + t_n = 1 \) by definitions of \( s \). Then \( sy + t_n x_n \) is a binary convex combination that we know is contained in \( C \) by definition.

**Proof of 417 (b).** Fix a cone \( C \). Let \( \sum_{k=1}^{n} t_k x_k \) be a conic combination of elements of \( C \). Each vector \( x_k \) lies on a closed ray, say \( r_k \), thus \( t_k x_k \) also lies on \( r_k \).

Therefore, we only need to show that the sum of two elements \( x_1, x_2 \in C \) is again in \( C \). This is true because \( x_1 + x_2 \) is a convex combination of \( 2x_1 \in r_1 \) and \( 2x_2 \in r_2 \).

**Proof of 417 (c).** Let \( X = \cap_{\alpha \in \mathcal{X}} X_{\alpha} \) be an intersection of convex sets. Take \( x, y \in X \) and \( t \in [0, 1] \). Then \( tx + (1 - t)y \in X_{\alpha} \) for all \( \alpha \in \mathcal{K} \), hence \( tx + (1 - t)y \in X \). Therefore, \( X \) is convex.

**Definition 418.** A \( k \)-simplex is the convex hull of \( k + 1 \) affinely independent vectors called the vertices of the simplex. The convex hull of any subset of the vertices is called a face of the simplex.

(a) A 0-simplex is a point.

(b) A 1-simplex is a line segment as defined in definition 416 (a).

(c) A 2-simplex is a triangle as defined in definition 426.

(d) A 3-simplex is called a tetrahedron.

**Definition 419.** A \( k \)-cell is a Cartesian product of \( k \) nonempty closed intervals of real numbers.

(a) A 0-cell is a point.

(b) A 1-cell is a closed interval.

(c) A 2-cell is called a rectangle. If a rectangle \( R \) is a product of two copies of the same interval, i.e. if \( R = [a, b]^2 \), we say that \( R \) is a square with side \( b - a \).

(d) A 3-cell is called a parallelepiped. If \( R = [a, b]^3 \), we say that \( R \) is a cube with side \( b - a \).
**Definition 420.** The following topology-independent definitions are often used for neighborhoods in a topological vector space $X$:

(a) $A$ is **absorbing** if $\bigcup_{k=0}^{\infty} kA = X$.
(b) $A$ is **symmetric** if $-A = A$.
(c) $A$ is **balanced** if $tA \subseteq A$ for any $t \in [0, 1]$.

**Definition 421.** The geometric version of linear independence has two special names: we say that the set $A \subseteq X$ of any vector space $X$ is **collinear** (on the same line) if $\dim(\text{span}(A)) \leq 1$ and **complanar** (on the same plane) if $\dim(\text{span}(A)) \leq 2$.

**Proposition 422.** Consider curve

$$\gamma : \mathbb{R} \to \mathbb{R}^n$$

$$\gamma(t) := (t, t^2, ..., t^n).$$

For any $t_1 < ... < t_n$, the points $\gamma(t_1), ..., \gamma(t_n)$ are linearly independent. This curve is called the **moment curve** of dimension $n$.

**Proof.** Follows from example 744. \qed
8.3. Analytic geometry in the plane

Remark 423. Analytic geometry is a XVII-century branch of mathematics that studies geometric figures using coordinate systems. The term “analytic geometry” may refer to a modern subbranch of algebraic geometry, however we refrain from using “analytic geometry” in that sense. Historically, most of these definitions were given either for the Euclidean plane or for the three-dimensional Euclidean space.

Most of the definitions from section 8.2 (Vector space geometry) are generalizations of concepts from analytic geometry. We will state definitions in the language of linear algebra and refrain from using synthetic (axiomatic) geometry. When working in the plane (resp. three-dimensional space), we will assume that we have fixed an orthonormal coordinate system $Oxy$ (resp. $Oxyz$), which allows us to visualize geometric figures.

**Definition 424.** Lines in $\mathbb{R}^2$ are, so ubiquitous that they can be represented by a lot of standard equations.

(a) When regarding a line as a parametric curve as in definition 407 (c), the formula

$$l(t) = tx + a$$

(137)

is called a **vector parametric equation**.

(b) Given eq. (137), the **scalar parametric equations** of the line are

$$\begin{cases} 
    l_1(t) = tx_1 + a_1 \\
    l_2(t) = tx_2 + a_2.
\end{cases}$$

(138)

(c) When regarding a line as an algebraic curve as in definition 407 (b), the equation

$$p(x, y) := Ax + By + C = 0$$

(139)

is called the **general equation** or simply **equation** of a line in a plane. Either $A$ or $B$ must be nonzero, so that $\deg(p) = 1$.

Note that multiple general equations can have the same locus (e.g. the entire polynomial ideal $\langle p \rangle$).

(d) If $A^2 + B^2 = 1$, we call eq. (139) a **normal equation**. This leaves us with only two representatives of $\langle p \rangle$.

(e) Given $k, m \in \mathbb{R}$ and $k \neq 0$, we define the **Cartesian equation** of a line:

$$y = kx + m.$$  

(140)

We call $k$ the **slope** of the line.

This is a special case of definition 424 (c) with $A = -k, B = -1$ and $C = m$. Unlike the general equation, the Cartesian equation of a line is unique.

Conversely, if $B \neq 0$ in eq. (139), we can define $k = -\frac{A}{B}$ and $m = -\frac{C}{B}$ to form a Cartesian equation.
Figure 13: A line in $\mathbb{R}^2$ defined using its Cartesian equation.

(f) Given nonzero $a, b \in \mathbb{R}$, we define the intercept equation of a line:

$$\frac{x}{a} + \frac{y}{b} = 1,$$

(141)

This is a special case of definition 424(c) with $A = \frac{1}{a}$, $B = \frac{1}{b}$ and $C = -1$. The intercept equation of a line is also unique.

If $A, B, C \neq 0$ in definition 424(c), we can define an intercept equation as $a = -\frac{C}{A}$ and $b = -\frac{C}{B}$.

**Definition 425.** A directed angle is a tuple of two closed rays with a common vertex. It is a closed cone. Given two rays $r_1, r_2$ with a common vertex, we denote their corresponding directed angle by $\alpha(r_1, r_2)$.

Suppose that $r_1$ and $r_2$ have scalar parametric equations

$$r_i : t \mapsto \begin{cases} t x_i + a_i \\ t y_i + b_i, \end{cases} \quad i = 1, 2.$$

We write

The condition of the rays having a common vertex is equivalent to $a_1 = a_2$ and $b_1 = b_2$. If not specified otherwise, we assume that $a_1 = a_2 = b_1 = b_2 = 0$.

The measure in radians of a directed angle, often called the angle itself, is defined as the number (see definition 430)

$$\alpha := \text{rem}(\arctan2(y_2, x_2) - \arctan2(y_1, x_1), 2\pi).$$

We can classify angles based on their measure as

(a) **zero** if $\alpha = 0$,

(b) **acute** if $\alpha \in (0, \frac{\pi}{2})$,

(c) **right** if $\alpha = \frac{\pi}{2}$,

(d) **obtuse** if $\alpha \in (\frac{\pi}{2}, \pi)$,

(e) **straight** if $\alpha = \pi$, in which case the angle is actually a line,

(f) **reflex** if $\alpha > \pi$. 

202
We often do not care about the order of the two rays and speak of an undirected angle. In this case, the measure of the undirected angle is the smaller of the measures of the two oriented angles. Thus, we cannot speak of straight and reflex undirected angles.

**Definition 426.** A triangle is a triple \((A, B, C)\) of points, no two of which are collinear (see definition 418 (c) for a more general definition). The three points are called the vertices of the triangle.

Define the associated line segments, called the sides of the triangle, and its (undirected) angles as

\[
\begin{align*}
\alpha & := \angle(b, c), \\
\beta & := \angle(a, c), \\
\gamma & := \angle(a, b).
\end{align*}
\]

Note that we defined the angles using segments rather than rays, but this is immaterial because each to each segment \([p, q]\) there corresponds exactly one closed ray \(t \mapsto p + tq\).

We can classify triangles based on their sides as

(a) **isosceles** if at least two of its sides have equal length
(b) **equilateral** if all of its sides have equal length

or based on their angles as

(a) **acute** if all of its angles are acute.
(b) **right** if at least one of the angles is straight.
(c) **obtuse** if at least one of its angles is obtuse.

**Definition 427.** The quadratic plane curves are algebraic curves given by a bivariate polynomial of degree 2. The general equation of a quadratic plane curve is

\[
c(x, y) := Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0.
\]  

(143)

Multiple equation can correspond to the same curve. Not all general equations, however, define algebraic curves. We will not concern ourselves with the details. See ?? ([UNDEFINED]) for a proof that the unit circle is an algebraic curve. It turns out that the algebraic curves given eq. (143) are precisely the ones listed here, collectively known as conic sections.

We give only canonical forms of the equations; any linear transformation of the corresponding loci is described by another general equation.
Figure 16: An ellipse, hyperbola and parabola defined via their parametric equations. The starting point is highlighted and the direction of the parametric curves is shown.

(a) An ellipse is a quadratic curve whose canonical equation has the form

\[ c(x, y) := \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0, \quad (144) \]

where \(a, b > 0\).

If \(a = b\), we say that the ellipse is a circle and we call \(a\) the circle’s radius. The unit circle is defined by \(a = b = 1\). Circles generalize to spheres in metric spaces. Definition 429 and definition 430 logically belong here, but are extracted separately for brevity.

We are often interested in defining ellipses via scalar parametric equations using trigonometric functions as follows:

\[
\begin{align*}
    x &= a \cos(t) \\
    y &= b \sin(t),
\end{align*}
\]

where \(t \in [0, 2\pi)\).

We will now demonstrate that eq. (144) and eq. (145) describe the same curve. First, suppose that the pair \((x_0, y_0)\) satisfies eq. (144). It follows from proposition 143 that \(t_0 := \arctan2\left(\frac{y_0}{x_0}, \frac{x_0}{a}\right)\) is a solution to the parametric equations. Conversely, if \(x_0 = a \cos(t_0)\) and \(y_0 = b \sin(t_0)\) for some \(t_0 \in [0, 2\pi)\), by proposition 136 (a) it follows that the pair \((x_0, y_0)\) is a root of eq. (144) and, by proposition 143, \(t_0\) can be restored given \(\cos(t_0)\) and \(\sin(t_0)\).

Therefore, every point of the parametric equation eq. (145) corresponds uniquely to a, solution of the canonical equation eq. (144) and vice versa, which makes the two approaches to defining ellipses equivalent.

(b) A hyperbola is a quadratic curve whose canonical equation has the form

\[ c(x, y) := \frac{x^2}{a^2} - \frac{y^2}{b^2} - 1 = 0, \quad (146) \]

where \(a, b > 0\).

Similarly to ellipses, we can define hyperbolas via scalar parametric equations using hyperbolic trigonometric functions as follows:

\[
\begin{align*}
    x &= a \cosh(t) \\
    y &= b \sinh(t),
\end{align*}
\]

where \(t \in \mathbb{R}\). This only defines the right part of the hyperbola. The left part is defined by replacing \(a\) with \(-a\).
(c) A **parabola** is a quadratic curve whose canonical equation has the form
\[
c(x, y) := y^2 - 2px = 0,
\]
where \( p \neq 0 \).

Unlike ellipses and hyperbolas, we do not define parametric equations. Instead, we define \( y \) as a function of \( x \) separately for the lower half-plane and upper half-plane:
\[
y(x) = \pm \sqrt{2px}.
\]

Ellipses, hyperbolas and parabolas are collectively called **conic sections**.

**Proposition 428. TODO:** Quadratic curve canonization

**Definition 429.** The definition of a circle of unit radius as the zero-locus of the polynomial \( x^2 + y^2 - 1 \) allows us to, solve a chicken-and-egg problem regarding the definitions of the number \( \pi \). It is conventional to define it as the ratio of a circle’s circumference to its diameter. For a unit circle, this diameter is 2. It will be simpler for us, however, to define \( \pi \) as the radius of a half-circle’s circumference since we can represent \( y \) as a function of \( x \) in the upper half-plane (see 17). Define the parametric curve
\[
y^+ : [-1, 1] \to [0, 1] \\
y^+(x) := \sqrt{1 - x^2}.
\]

We use corollary 84 to find the length of the graph \( gph(y^+(x)) \). The derivative of \( y^+(x) \) is
\[
D_x[y^+(x)] = \frac{-2x}{2\sqrt{1 - x^2}} = -\frac{x}{\sqrt{1 - x^2}} dx.
\]

The length of the curve \( gph(y^+) \) is thus
\[
\text{len}(gph(y^+)) = \int_{-1}^{1} \sqrt{1 + \frac{x^2}{1 - x^2}} dx = \int_{-1}^{1} 1 \sqrt{1 - x^2} dx.
\]

This justifies the definition
\[
\pi := \int_{-1}^{1} \frac{1}{\sqrt{1 - x^2}} dx.
\]

See lemma 137 for a proof of how this relates to the trigonometric functions and proposition 145 (a) as a consequence.
Figure 18: An “abstract” right triangle in the Euclidean plane with legends for geometric sines and cosines and the same triangle in $\mathbb{R}^2$ connecting the origin to a point $(x_0, y_0)$ on the unit circle.

**Definition 430.** After defining the trigonometric functions $\cos(z)$ and $\sin(z)$ analytically via power series, we will define their geometric counterparts $\cos_G(z)$ and $\sin_G(z)$ and show the connection between them. The actual geometric definition relies on formalisms that are far beyond our interest (see the notes in definition 395).

Fix a point $(x_0, y_0)$ on the unit circle (that is, $x_0^2 + y_0^2 = 1$) and define the points

$$A := (x_0, y_0),$$
$$B := (0, 0),$$
$$C := (x_0, 0).$$

Consider the triangle formed by these vertices. Figure 18 illustrates the situation. The original “geometric definition” of $\sin_G$ and $\cos_G$ regards them as functions of an angle rather than numeric functions. $\sin_G$ and $\cos_G$ are only defined for two of the angles in a right triangle. The geometric definition is

$$\sin_G(\alpha) := \frac{\text{len}(b)}{\text{len}(c)}, \quad \cos_G(\alpha) := \frac{\text{len}(a)}{\text{len}(c)},$$
$$\sin_G(\beta) := \frac{\text{len}(b)}{\text{len}(c)}, \quad \cos_G(\beta) := \frac{\text{len}(a)}{\text{len}(c)}.$$

In our case, $\text{len}(a) = y_0$, $\text{len}(b) = x_0$ and $\text{len}(c) = 1$. Furthermore, $\sin_G(\beta)$ nor $\cos_G(\beta)$ are immaterial to our subsequent arguments and we only introduced them for the sake of having a full definition.

Therefore, we conclude that

$$\sin_G(\alpha) = x_0, \quad \cos_G(\alpha) = y_0.$$

To see that $\sin_G$ and $\cos_G$ are somewhat analogous to $\sin$ and $\cos$, notice that by proposition 143, there exists a unique $t_0 := \arctan2(y_0, x_0)$ such that

$$\sin(t_0) = x_0, \quad \cos(t_0) = y_0.$$

Therefore, our analytic definition of the trigonometric functions as numeric functions correspond to the classical geometric definition in the special case where we consider the angle near the origin in the triangle formed by the vertices eq. (151). This motivates “measuring” angles using the obtained correspondence. This unit of measurement is called a **radian**. We say that the angle $\alpha$ is $t_0$ radians. Outside of mathematics, it is more conventional to use **degrees**, which are obtained from radians by scaling with $\frac{180}{\pi}$. That is, $\alpha$ is $\frac{180}{\pi} t_0$ degrees.
8.4. Manifolds

**Definition 431.** Let $X$ be a topological space and $Y$ be a topological vector space.

A **coordinate chart** on $X$ over $Y$ is a pair $(U_\alpha, \varphi_\alpha)$, where

- $U_\alpha \subseteq X$ is a connected open set.
- $\varphi_\alpha : U_\alpha \to Y$ is a homeomorphic embedding, called a **coordinate homeomorphism**.

An **atlas** on $X$ over $Y$ is an indexed family $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in \mathcal{K}}$ of charts such that the family $\{U_\alpha\}_{\alpha \in \mathcal{K}}$ is a cover of $X$. If $Y = \mathbb{K}^n$ for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, we say that $X$ is a real (resp. complex) manifold of dimension $n$.

For any two coordinate homeomorphisms $\varphi_\alpha$ and $\varphi_\beta$ in an atlas, the function restriction of the composition $\varphi_\alpha \circ \varphi^{-1}_\alpha$ to $U_\alpha \cap U_\beta$ is a homeomorphism from $\varphi_\alpha(U_\alpha \cap U_\beta)$ to $\varphi_\beta(U_\alpha \cap U_\beta)$, called a **transition map**.

**Definition 432.** We call the topological space $X$ a **topological manifold** if it has a countable atlas. If $Y = \mathbb{R}^n$, we say that $X$ is a manifold of dimension $n$.

**Definition 433.** We call the topological manifold $X$ a **smooth manifold** of type $C^k$ if the transition maps are $k$-times continuously Fréchet differentiable.

We also allow $k = \infty$ for infinitely differentiable transition maps $k = \omega$ for analytic transition maps.
8.5. Quadratic curves

This subsection is dedicated to curves in $\mathbb{R}^2$ described by quadratic polynomials. We first formalize the meaning of algebraic variety and algebraic curve, but do not go into algebraic geometry beyond that. After that, we restrict ourselves to the classical theory of quadratic curves.

Definition 434. The Krull dimension $\text{dim} R$ of a commutative unital ring $R$ is the ordinal supremum of the lengths $n$ of chains

$$\{0\} \subseteq P_1 \subseteq P_2 \subseteq \cdots \subseteq P_n \subseteq R$$

of nontrivial prime ideals strictly contained in $R$.

The Krull dimension is either a nonnegative integer or the smallest limit cardinal $\aleph_0$, which we denote via $\infty$.

Proposition 435. Krull dimensions have the following properties:

(a) The dimension of a field is 0.

(b) If $I$ is an ideal of $R$, then $\text{dim } R/I \leq \text{dim } R$.

(c) We have $1 \leq \text{dim } R \leq n$ if and only if $\text{dim } S^{-1} R \leq n - 1$ for every $x \in R$, where

$$S_{[x]} = \{z \in R \mid \exists n \in \mathbb{Z}_{\geq 0} : \exists a \in R : z = x^n(1 + ax)\}.$$  

(d) We have $1 \leq \text{dim } R \leq n$ if and only if, for every sequence $x_1, \ldots, x_{n+1}$ in $R$, there exists a sequence $a_1, \ldots, a_{n+1}$ in $R$ and nonnegative integers $m_1, \ldots, m_{n+1}$ such that

$$x_1^{m_1}(\cdots (x_{n+1}^{m_{n+1}}(1 + a_{n+1}x_{n+1}) + \cdots ) + a_1x_1) = 0.$$  

(e) For an algebra $M$ over a field $\mathbb{K}$, if every sequence $x_1, \ldots, x_{n+1}$ is algebraically dependent, then $\text{dim } M \leq n$.

(f) For a field $\mathbb{K}$, the polynomial ring $\mathbb{K}[X_1, \ldots, X_n]$ has dimension $n$. We call $\mathbb{K}[X_1, \ldots, X_n]$ the $n$-dimensional affine space over $\mathbb{K}$.

Proof.

Proof of 435 (a). A field is a simple ring by definition, hence it has zero proper nontrivial ideals.

Proof of 435 (b). Follows from theorem 651 (Quotient ideal lattice theorem).

Proof of 435 (c). Suppose that

$$\langle 0_R \rangle \subseteq S^{-1} P_1 \subseteq \cdots \subseteq S^{-1} P_m \subseteq S^{-1} R$$

is a chain of prime ideals in $S^{-1} R$. By proposition 596 (b),

$$\langle 0_R \rangle \subseteq P_1 \subseteq \cdots \subseteq P_m \subseteq R$$

is a chain of prime ideals in $R$. Then clearly $m \leq \text{dim } R$.

We will now show that $m \leq \text{dim } R - 1$. We will need the following observations:

PROOF OF 435 (c).
• For every \( x \in R \), every maximal ideal \( M \) in \( R \) intersects \( S_{\{x\}} \). This is obvious if \( x \in M \). Otherwise, since \( M \) is maximal, \( R = M + \langle x \rangle \). Hence, there exist \( m \in M \) and \( a \in R \) such that \( m + ax = 1 \). Then \( 1 - ax \in M \cap S_{\{x\}} \).

• No nested prime ideal \( P \subset M \) intersects \( S_{\{x\}} \) for \( x \in M \setminus R \). Indeed, suppose that \( x^n(1 + ax) \in S_{\{x\}} \cap P \). Since \( P \) is prime, \( x^n \not\in P \), and hence \( 1 + ax \in P \). Since \( x \in M \), then \( ax \in M \), and thus from \( 1 + ax \in M \) it follows that \( 1 \in M \), contradicting the maximality of \( M \).

It follows that for any maximal ideal \( M \) of \( R, S^{-1}_{\{x\}} \) is not a prime ideal for any \( x \in R \), while for any prime ideal \( P \subset M \) and \( x \in M \setminus P, S^{-1}_{\{x\}} \) is a prime ideal. Therefore, \( m \leq \dim R - 1 \) and, in general, this estimate is strict.

**Proof of 435 (d).**

**Proof of sufficiency.** We use induction on \( n \). The base case \( n = 1 \) follows directly from proposition 435 (c).

Suppose that the statement holds for \( n - 1 \) and let \( \dim R = n \). By proposition 435 (c), \( \dim S^{-1}_{\{x\}} R \leq n - 1 \) for every \( x \in R \).

Let \( x_1, \ldots, x_{n+1} \) be members of \( R \). By using the inductive hypothesis on \( S^{-1}_{\{x_{n+1}\}} R \), we obtain members \( a_1, \ldots, a_n \) and \( b_1, \ldots, b_n \) of \( R \) and nonnegative integers \( m_1, \ldots, m_n \) and \( k_1, \ldots, k_n \), such that

\[
x_1^{m_1} \left( \cdots \left( x_n^{m_n} \left( 1 + \frac{a_n}{x_{n+1}^{k_n} (1 + b_n x_{n+1})} x_n \right) + \cdots \right) + \frac{a_1}{x_{n+1}^{k_1} (1 + b_1 x_{n+1})} x_1 \right) = 0.
\]

Put

\[
c_1 := \prod_{j \neq i} x_{n+1}^{k_j} (1 + b_j x_{n+1})
\]

and

\[
c := \prod_{j=1}^{n+1} x_{n+1}^{k_j} (1 + b_j x_{n+1}).
\]

Let \( m_{n+1} := \prod_{j=1}^{n+1} k_j \). Since \( c \) is a member of \( S_{\{x_{n+1}\}} \), there exists \( a_{n+1} \) (which we can obtain explicitly if desired) such that

\[
c = x_{n+1}^{m_{n+1}} (1 + a_{n+1} x_{n+1}).
\]

Then

\[
x_1^{m_1} \left( \cdots \left( x_n^{m_n} \left( x_{n+1}^{m_{n+1}} (1 + a_{n+1} x_{n+1}) + a_n c_n x_n \right) + \cdots \right) + a_1 c_1 x_1 \right) = 0.
\]

**Proof of necessity.** Follows from following the sufficiency subproof in the reverse direction.
Proof of 435 (e). Suppose that every sequence of \( n + 1 \) members of \( M \) is algebraically dependent. Let \( x_1, \ldots, x_{n+1} \) be members of \( M \). Then there exists a polynomial \( p(x_1, \ldots, x_{n+1}) \) such that \( p(x_1, \ldots, x_{n+1}) = 0 \).

The polynomial can be written (in multiple ways) as

\[
p(X_1, \ldots, X_{n+1}) = X_1^{\gamma_1}\left(\cdots (X_{n+1}^{\gamma_{n+1}}(1 + q_{n+1}(X_{n+1})X_{n+1}) + \cdots) + q_1(X_1, \ldots, X_{n+1})X_1\right).
\]

Then, by proposition 435 (d), \( \dim M \leq n \).

Proof of 435 (f). The vector space \( \mathbb{K}[X_1, \ldots, X_n] \) has dimension \( n \). By proposition 661 (c), every \( n + 1 \) polynomials are algebraically independent. From proposition 435 (e) it follows that \( \dim \mathbb{K}[X_1, \ldots, X_n] \leq n \).

Furthermore, we have the following chain of prime ideals

\[
\langle 0 \rangle \subsetneq \langle X_1 \rangle \subsetneq \langle X_1, X_2 \rangle \subsetneq \langle X_1, \ldots, X_n \rangle \subsetneq \mathbb{K}[X_1, \ldots, X_n].
\]

Therefore, \( \dim \mathbb{K}[X_1, \ldots, X_n] \geq n \). \( \square \)

**Definition 436.** For each ideal \( I \) of the polynomial ring \( \mathbb{K}[X_1, \ldots, X_n] \) over a field \( \mathbb{K} \), we define its **affine algebraic set**\(^3\) as the locus of the simultaneous roots of all polynomials in \( I \),

\[
\mathcal{V}(I) := \{(x_1, \ldots, x_{n+1}) \in \mathbb{K}^n \mid \forall p \in I. p(x_1, \ldots, x_{n+1}) = 0\}.
\]

(a) The quotient \( \mathbb{K}[X_1, \ldots, X_n]/I \) is called a **coordinate ring** of \( \mathcal{V}(I) \).

(b) The **dimension** \( \dim(\mathcal{V}(I)) \) of an affine set is defined as the **Krull dimension** of the coordinate ring \( \mathbb{K}[X_1, \ldots, X_n]/I \).

By theorem 607 (Quotient submodule lattice theorem), the dimension is the supremum of the length of prime ideal chains starting at \( I \):

\[
I \subsetneq P_1 \subsetneq P_2 \subsetneq \cdots \subsetneq P_n \subsetneq \mathbb{K}[X_1, \ldots, X_n].
\]

(c) If \( I \) is a prime ideal, we say that \( \mathcal{V}(I) \) is an **algebraic variety**.

(d) An **algebraic curve** over \( \mathbb{K}^n \) is an affine variety of dimension one.

(e) An **algebraic surface** over \( \mathbb{K}^n \) is an affine variety of dimension two.

**Proposition 437.** We will work in the ring \( \mathbb{R}[X, Y] \) of real polynomials in two indeterminates. Take the quadratic polynomial

\[
p(X, Y) := aX^2 + bXY + cY^2 + dX + eY + f.
\]

Then the affine algebraic set \( \mathcal{V}((p(X,Y))) \) is an **algebraic curve** if and only if \( p(X,Y) \) is an irreducible polynomial.

\(^3\) bg: алгебраично многообразие, ru: алгебрическое многообразие

---

210
Proof. By proposition 691, $\mathbb{R}[X, Y]$ is a GCD domain, and, by proposition 674 (c), every irreducible polynomial is prime.

Thus, if $p(X, Y)$ is a nonconstant irreducible polynomial, then the following Hasse diagram shows how the principal ideal of $p(X, Y)$ relates to other prime ideals

\[
\begin{array}{c}
\mathbb{R}[X, Y] \\
\langle X, Y \rangle \\
\langle X \rangle \\
\langle p(X, Y) \rangle \\
\{0\}
\end{array}
\]

By proposition 435 (f), the Krull dimension of $\mathbb{R}[X, Y]$ is 2. Hence, by theorem 651 (Quotient ideal lattice theorem), the chain of quotients

\[
\frac{\langle p(X, Y) \rangle}{\langle p(X, Y) \rangle} \subseteq \frac{\langle X, Y \rangle}{\langle p(X, Y) \rangle} \subseteq \frac{\mathbb{R}[X, Y]}{\langle p(X, Y) \rangle}
\]

is a maximal chain of prime ideals.

Therefore, the coordinate ring of $\mathcal{V}(\langle p(X, Y) \rangle)$ has Krull dimension 1, and hence the affine algebraic set itself is an algebraic curve. \qed

Lemma 438. Fix an integral domain $D$ and let $a$ and $b$ be some members of $D$. Assume that $a$ is a unit. Then the polynomial $p(X)$ is irreducible over $D[X]$ if and only if $p(aY + b)$ is irreducible over $D[Y]$.

Proof. Consider the evaluation $X \mapsto aY + b$ and the corresponding evaluation homomorphism $\Phi : D[X] \rightarrow D[Y]$.

Since $a$ is a unit, we can also consider the evaluation homomorphism $\Psi : D[Y] \rightarrow D[X]$ given by

\[
Y \mapsto \frac{X}{b} + \frac{a}{b}.
\]

Clearly $\Phi$ and $\Psi$ are mutual inverses, hence both are isomorphisms. By proposition 666 (d), $p(X)$ is irreducible in $D[X]$ if and only if $\Phi(p) = p(aY + b)$ is irreducible in $D[Y]$. \qed

211
9. Group theory

Modern algebra takes its roots in abstracting integers and real numbers and their addition and multiplication. Both of these operations are commutative and, if we want to generalize their properties, it is sensible to study commutative operations.

Another type of objects that usually fits in the same algebraic framework are functions and their composition. Functions from a set to itself can be composed to form another function of the same type, similarly to how two integers can be added to obtain another integer. The main difference is in the non-commutativity of function composition.

This suggests that we use the same algebraic structures to study both generalizations of numbers and generalizations of functions over a set. The first case is commutative, the second is not. This is why commutative and non-commutative structures, even though they are similarly defined, can have very different properties and applications.

We shall not attempt to give a precise definition for an algebraic structure. There are very general frameworks for doing so, however their complexity is unjustified for us. We will instead build standard algebraic structures from “base building blocks”, although we will utilize very general definitions like categorical kernels and images defined in definition 454.

The simplest algebraic structures that will be of interest to us are groups and their less well-behaved generalizations, semigroups and monoids.

Except as a building block for more complicated algebraic structures, groups arise whenever some mathematical structure exhibits symmetries, and this concept is formalized via automorphism groups and group actions. If, instead of symmetries we have non-invertible but nonetheless well-behaved transformations, we can instead study endomorphism monoids and monoid actions.
9.1. Monoids

We list here several basic algebraic structures, we will use mostly as building blocks for more complicated structures.

As discussed in remark 834, listing all operations explicitly is cumbersome, and we will usually avoid it.

**Definition 439.** The simplest algebraic structure is a **pointed set**. It is simply a nonempty set \( X \) equipped with a distinguished element \( e \). It is an algebraic structure because \( e \) can be regarded as the sole value of a nullary function *: \( X^0 \to X \).

We will call \( e \) the **origin** of \( X \) based on the terminology for affine coordinate systems.

Pointed sets have the following metamathematical properties:

(a) Pointed sets can also be viewed as models of an empty theory for a first-order logic language with a constant symbol, i.e. a nullary functional symbol.

(b) A homomorphism between the pointed sets \((X,e_X)\) and \((Y,e_Y)\) is, explicitly, a function \( \varphi : X \to Y \) that satisfies

\[
\varphi(e_X) = e_Y.
\]

(c) The set \( S \subseteq X \) is a submodel of \( X \) if \( e \in S \).

In particular, as a consequence of proposition 863, the image of a pointed set homomorphism is a submodel of its range.

(d) The trivial pointed set is, up to an isomorphism, the set \( \{e\} \).

It is a zero object in \( \textbf{Set} \), as discussed in example 1131 (b).

(e) We denote the category of \( \mathcal{U} \)-small models for this theory by \( \mathcal{U} \text{-}\textbf{Set}_* \).

**Definition 440.** A **set with an involution** is a set \( X \) with a unary operation \((\cdot)^{-1}\) such that

\[
(x^{-1})^{-1} = x
\]

for every \( x \in X \).

Such an operation is called, surprisingly, an **involution**.

Sets with involutions have the following metamathematical properties:

(a) We define the theory of sets with involution as a theory over the language consisting of a single unary functional symbol \( \cdot^{-1} \) with the sole axiom

\[
(\xi^{-1})^{-1} = \xi.
\]

(b) A homomorphism between sets with involutions \( X \) and \( Y \) is a function \( \varphi : X \to Y \) satisfying

\[
\varphi(x^{-1}) = \varphi(x)^{-1}.
\]
(c) Any subset of a set with involution is again a set with involution.
In particular, as a consequence of proposition 863, the image of a homomorphism
\( \varphi : X \to Y \) is a submodel of \( Y \).

(d) The trivial set with involution is the empty set.

(e) We denote the category of \( \mathcal{U} \)-small models for this theory by \( \mathcal{U} \text{-Inv} \).

Definition 441. A magma is a set \( M \) equipped with a binary function \( \cdot : M \times M \to M \),
called the magma operation. Unless specified otherwise, we denote this operation by
juxtaposition as \( x \cdot y \) instead of \( x \cdot y \).

We often call the operation multiplication or, in the case of endomorphism monoids —
composition. See also the notes in remark 512 regarding additive magmas and in definition 1177 regarding the order of operands.

(a) In analogy to the theory of pointed sets, we can define the theory of magmas as an
empty theory over a language with a single infix binary functional symbol.

(b) A homomorphism between the magmas \( (M, \cdot_M) \) and \( (N, \cdot_N) \) is, explicitly, a function
\( \varphi : M \to N \) such that
\[ \varphi(x \cdot_M y) = \varphi(x) \cdot_N \varphi(y) \]  
for all \( x, y \in M \).

(c) The set \( A \subseteq M \) is a first-order submodel of \( M \) if it is closed under the magma operation.
That is, if \( x, y \in A \) implies \( xy \in A \).

We call \( A \) a submagma of \( M \).

As a consequence of proposition 863, the image of a magma homomorphism is a
submagma of its range.

(d) The trivial magma is the empty set with an empty operation. It is the unique zero
object in \( \text{Mag} \).

(e) We define an additional exponentiation operation for positive integers \( n \) recursively
as
\[ x^n := \begin{cases} x, & n = 1 \\ x^{n-1} \cdot x, & n > 1 \end{cases} \]  
(156)

(f) It is customary to perform magma operations with sets. That is, if \( A \) and \( B \) are sets
in the magma \( M \), it is customary to write
\[ A \cdot B := \{ a \cdot b : a \in A, b \in B \}. \]

This actually turns the power set \( \text{pow}(M) \) into a magma, which we will call the power
set magma of \( M \). This is especially useful with the convention remark 930 since it
allows us to write \( aB \) for \( a \in M \) and \( B \subseteq M \).

See proposition 443.

Note that this concept is not exactly the same as that of product ideals.
We denote the category of $\mathcal{U}$-small models for the theory of magmas by $\mathcal{U}$-$\text{Mag}$.

The opposite magma of $(M, \cdot)$ is the magma $(M, \star)$ with multiplication reversed:

$$x \star y := y \cdot x.$$ 

We denote the opposite magma by $M^{\text{opp}}$. This is justified in definition 441 (h).

We list some additional restrictions that are often imposed on magmas.

(i) We can add the (universal closure of) following axiom to the theory:

$$(\xi \cdot \eta) \cdot \zeta = \xi \cdot (\eta \cdot \zeta). \quad (157)$$

If (157) is satisfied, we say that the operation $\cdot$ and, by extension, the magma itself, are associative. Associative magmas are usually called semigroups. Associativity imposes no additional restrictions on the homomorphisms, hence semigroups are a full subcategory of $\text{Mag}$.

(j) Another common axiom is commutativity:

$$\xi \cdot \eta \doteq \eta \cdot \xi. \quad (158)$$

Commutative magmas also form a full subcategory. Obviously $M = M^{-1}$ in a commutative magma.

(k) We say that the operation $\cdot$ is idempotent if

$$\xi \cdot \xi \doteq \xi. \quad (159)$$

(l) We say that $\cdot$ is left-cancellative if

$$\forall \zeta. (\zeta \cdot \xi \doteq \zeta \cdot \eta) \rightarrow \xi = \eta \quad (160)$$

and right-cancellative if

$$\forall \zeta. (\xi \cdot \zeta \doteq \eta \cdot \zeta) \rightarrow \xi = \eta \quad (161)$$

The operation is cancellative if it is both left and right cancellative. Cancellative magmas also form a full subcategory.

Example 442. We list several examples of magmas satisfying different properties.

(a) Associative binary operations on a set are abundant and are part of the definition of essential algebraic structures like groups, semirings, semimodules and (semi)lattices. These operations are homogeneous in the sense that their signature only contains a single set, unlike group actions and scalar products in semimodules.
(b) The quintessential example of a non-commutative operation is composition in any set of functions or, more generally, morphism composition in any category. Composition is associative. Cancellation with respect to composition is discussed in definition 1125 and, for function composition, in proposition 989.

(c) The midpoint operation

\[(x, y) \mapsto \frac{x + y}{2}\]

makes \(\mathbb{R}\) a commutative and cancellative magma, which is not associative.

**Proposition 443.** Associativity and commutativity from a magma \(M\) are preserved in \(\text{pow}(M)\), unlike cancellation.

**Proof.** Associativity and commutativity are obviously preserved.

To show that cancellation is not, consider the group \(F_2\). It is a cancellative magma by proposition 457 (a). Define the sets \(A := \{0, 1\}\) and \(B := \{0\}\). Then

\[A + A = A = A + B,\]

however we cannot cancel \(A\) from the left because \(A \neq B\). □

**Proposition 444.** Fix a magma \(M\). Magma exponentiation in \(M\) has the following basic properties:

(a) We have the following commutativity-like property: for \(x \in M\) and \(n = 1, 2, \ldots\),

\[x^n = xx^{n-1} = x^{n-1}x.\] (162)

(b) Exponentiation distributes over multiplication: for any member \(x \in M\) and any two positive integers \(n\) and \(m\),

\[x^{n+m} = x^n x^m.\] (163)

(c) For any member \(x \in M\) and any two positive integers \(n\) and \(m\),

\[(x^n)^m = x^{nm}.\] (164)

**Proof.**

**Proof of 444 (a).** We use induction on \(n\). The cases \(n = 1\) and \(n = 2\) are obvious. For \(n > 2\), we have

\[x^n \overset{(156)}{=} xx^{n-1} \overset{\text{ind}}{=} xx^{n-2} x \overset{(156)}{=} x^{n-1}x.\]

**Proof of 444 (b).** We use induction on \(n\). The case \(n = 1\) follows directly from (156). The case \(n > 1\) follows from

\[x^{n+m} \overset{(156)}{=} xx^{n+(m-1)} \overset{\text{ind}}{=} xx^{n-1}x^m \overset{(156)}{=} x^n x^m.\]
Proof of 444 (c). We use induction on $n$. The case $n = 1$ is obvious and the rest follows from

\[(x^n)^m \overset{(156)}{=} x^n(x^n)^{m-1} \overset{\text{ind.}}{=} x^n x^n (m-1) \overset{(163)}{=} x^{nm}.\]

\[\square\]

Definition 445. An ordered magma\footnote{Based on [Gol10, p. 224]} is a commutative magma $M$ equipped with a partial order $\leq$ such that $x \leq y$ implies $xz \leq yz$ for every $z \in M$.

The condition for commutativity is not necessary, and the partial order can be generalized to a preorder, but neither generalization will be of any use for us.

The category of small ordered magmas is a concrete category over both Mag and Pos.

Example 446. We list several examples of ordered magmas.

(a) The natural numbers with addition form an ordered magma as a consequence of proposition 11; and so do $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$.

(b) More generally, every limit ordinal as the set of all smaller ordinals is an ordered magma under ordinal addition.

Unlike addition in natural numbers, however, ordinal addition is not commutative as shown in example 1074.

(c) Every join-semilattice $(X, \vee)$ is an ordered magma with the lattice order. Indeed, if $x \leq y$, then

- If $z \leq x$, then $x = x \vee z \leq y \vee z = y$.

- If $x \leq z \leq y$, then $z = x \vee z \leq y \vee z = y$.

- If $z \geq y$, then $z = x \vee z \leq y \vee z = z$.

Every meet-semilattice is also an ordered magma.

Definition 447. A monoid is an associative magma with a distinguished element $e$ such that $ex = x = xe$ for every element $x$. Such an element is obviously unique, and we call it the identity or neutral element of the monoid. This makes monoids pointed sets.

The requirement of associativity is conventional but not strictly necessary. Non-associative monoids will not be useful to us.

(a) The theory of monoids consists of associativity and the axiom

\[
\forall \xi. (e \cdot \xi \equiv \xi \land \xi \cdot e \equiv \xi) \quad (165)
\]

over the combined language of pointed sets and magmas.
(b) A homomorphism between monoids is a function that satisfies both (152) and (155).

(c) The set $A \subseteq X$ is a submodel if $X$ if $e \in A$. This is equivalent to $A$ being a pointed subset.

We say that $A$ is a submonoid.

As a consequence of proposition 863, the image of a monoid homomorphism is a submonoid of its range.

(d) The trivial monoid is the trivial pointed set $\{e\}$.

(e) We extend magma exponentiation to all nonnegative integers by defining

$$x^0 := e.$$  

We denote the subcategory of commutative monoids by $\mathcal{U}$-$\text{CMon}$. We usually write commutative monoids additively as explained in remark 512.

(f) The power set magma $\text{pow}(M)$ of a monoid $M$ with identity $e$ is again a monoid with identity $\{e\}$.

(g) The category of $\mathcal{U}$-small models $\mathcal{U}$-$\text{Mon}$ of monoids is concrete with respect to both $\mathcal{U}$-$\text{Set}$, and $\mathcal{U}$-$\text{Mag}$.

We will discuss the free monoid functor in theorem 471 (Free monoid universal property). Then from proposition 917 it will follow that monomorphisms are injective functions.

For epimorphisms a similar statement does not hold, unfortunately. The embedding $\iota: \mathbb{N} \to \mathbb{Z}$ is an epimorphism by theorem 521 (Grothendieck monoid completion universal property), but it is clearly not surjective.

(h) The delooping of the opposite magma for a monoid $M$ is the opposite category of the delooping of $M$. This justifies the notation $M^{\text{op}}$

Example 448. We list several important examples of monoids.

(a) The nonnegative natural numbers with addition form a quintessential example of a monoid. We prove in proposition 6 that they are a monoid.

(b) Another important example of a monoid is the Kleene star $\mathcal{A}$ over some alphabet $\mathcal{A}$.

The importance for monoid theory comes from the free monoid universal property described in theorem 471 (Free monoid universal property).

(c) Every bounded join-semilattice is a monoid as a consequence of (511), and similarly for meet-semilattice.
[Pin14] **Example 449.** Monoid homomorphisms may not preserve the cancellation property. For example, the natural numbers \( \mathbb{N} \) are a cancellative monoid under addition, as shown in proposition 6, but the homomorphism

\[
h : (\mathbb{N}, +) \to (F_2, \max)
\]

\[
h(n) :=
\begin{cases}
0, & n = 0 \\
1, & n > 0
\end{cases}
\]

does not preserve the cancellative property. Indeed, \( \max\{0, 1\} = \max\{1, 1\} \), but \( 0 \neq 1 \).

**Definition 450.** The **direct product** of a family of monoids \( \{M_k\}_{k \in \mathcal{K}} \) is their Cartesian product \( \prod_{k \in \mathcal{K}} M_k \) with the componentwise operation

\[
\{x_k\}_{k \in \mathcal{K}} \cdot \{y_k\}_{k \in \mathcal{K}} := \{x_k \cdot y_k\}_{k \in \mathcal{K}}.
\]

For every index \( m \in \mathcal{K} \), we define the canonical projection

\[
\pi_m : \prod_{k \in \mathcal{K}} M_k \to M_m
\]

\[
\pi_m(\{x_k\}_{k \in \mathcal{K}}) := x_m.
\]

If all \( M_k \) are equal to \( M \), we denote the direct product by \( M^\mathcal{K} \). The **direct sum** \( \bigoplus_{k \in \mathcal{K}} M_k \) is the submonoid of the direct product, in which only finitely many components of each tuple are distinct from the identity.

For every index \( m \in \mathcal{K} \), we define the canonical embedding

\[
t_m : M_m \to \bigoplus_{k \in \mathcal{K}} M_k
\]

\[
t_m(x_m) :=
\begin{cases}
x_m, & k = m \\
e_k, & k \neq m
\end{cases}
\]

If all \( M_k \) are equal to \( M \), we denote the direct sum by \( M^{\oplus \mathcal{K}} \).

Unlike the product, the direct sum is mostly useful for commutative monoids. The role of direct sums is discussed in remark 794 (c iii) and, more concretely, in proposition 451.

**Proposition 451.**

(a) The categorical product of the family \( \{M_k\}_{k \in \mathcal{K}} \) in the category \textbf{Mon} of monoids is their direct product \( \prod_{k \in \mathcal{K}} M_k \).

(b) The categorical coproduct of the family \( \{M_k\}_{k \in \mathcal{K}} \) in the category \textbf{CMon} of commutative monoids is their direct sum \( \bigoplus_{k \in \mathcal{K}} M_k \).

Compare this to the non-commutative case for groups discussed in proposition 486 (b).

**Proof.**
**Proof of 451 (a).** Let \((A, \alpha)\) be a cone for the discrete diagram \(\{M_k\}_{k \in \mathcal{K}}\). We want to define a monoid homomorphism \(l_A : A \to \prod_{k \in \mathcal{K}} M_k\) such that, for every \(m \in \mathcal{X}\) and \(a \in A\),

\[
\alpha_m(a) = \pi_m(l_A(a)).
\]

This suggests the definition

\[
l_A(a) := \{\alpha_k(a)\}_{k \in \mathcal{K}}.
\]

**Proof of 451 (b).** Let \((A, \alpha)\) be a cocone for the discrete diagram \(\{M_k\}_{k \in \mathcal{K}}\). We want to define a monoid homomorphism \(l_A : \bigoplus_{k \in \mathcal{K}} M_k \to A\) such that, for every \(m \in \mathcal{K}\) and \(x \in M_m\),

\[
\alpha_m(x) = l_A(u_m(x)).
\]

This suggests the definition

\[
l_A(\{x_k\}_{k \in \mathcal{K}}) := \prod_{k \in \mathcal{K}} \alpha_k(x_k).
\]

We discuss well-definedness of infinitary operations in direct sums in remark 794 (c iii). \(\square\)
9.2. Groups

**Definition 452.** Let $M$ be a monoid. We say that $y$ is the left inverse (resp. right inverse) of $x$ if $yx = e$ (resp. $xy = e$).

If $y$ is simultaneously a left and right inverse of $x$, we call a two-sided inverse or simply an inverse of $x$ and denote it by $x^{-1}$. It is unique by proposition 453. This notation is consistent with monoid exponentiation defined in definition 447(e).

We call $x$ invertible if it has a two-sided inverse.

**Proposition 453.** For every element $x$ of any monoid, the (two-sided) inverse $x^{-1}$ of $x$, if it exists, is unique.

**Proof.** If $y$ and $z$ are both inverses of $x$, then

$$y = ey = zxy = ze = z.$$

\[\square\]

**Definition 454.** Let $C$ be a pointed category with a fixed zero object $Z$.

(a) For every pair of objects $A$ and $B$ in $C$, there exists unique morphism, called the zero morphism, that uniquely factors through $Z$:

$$A \longrightarrow Z \longrightarrow B$$

We denote this zero morphism by $0_{A,B}$.

(b) The kernel cone of a morphism $f : A \rightarrow B$ is the equalizer cone of $f$ and $0_{A,B}$.

By proposition 1216, a kernel morphism is necessarily a monomorphism. A monomorphism is normal if it is a kernel.

(c) Dually, the cokernel cocone of a morphism $f : A \rightarrow B$ is the coequalizer cocone of $f$ and $0_{A,B}$.

By proposition 1216, a cokernel morphism is necessarily an epimorphism. An epimorphism is normal if it is a kernel.

**Definition 455.** A group is a monoid in which every element has an inverse. Groups are the most well-studied and most well-behaved magmas. Many useful properties like cancellation rely on associativity, so we do not consider non-associative groups.

Groups have the following metamathematical properties:

(a) We can construct a first-order theory for groups by adding a unary functional symbol $(\cdot)^{-1}$ to the language and the axiom

$$\forall \xi . (\xi \cdot \xi^{-1} = e \land \xi^{-1} \cdot \xi = e)$$

(167)

to the theory of monoids.
(b) A function \( \varphi : G \to H \) between two groups is called **even** if, for every \( x \in G \), we have

\[
\varphi(x^{-1}) = \varphi(x)
\]

(168)

and **odd** if

\[
\varphi(x^{-1}) = \varphi(x)^{-1}.
\]

(169)

(c) A first-order homomorphism between the groups \( G \) and \( H \) is an odd monoid homomorphism.

As shown in proposition 458, however, the conditions (152) and (169) are redundant.

(d) The set \( A \subseteq G \) is a submodel of \( G \) if it is a submonoid and if \( x \in A \) implies \( x^{-1} \in A \).

We say that \( A \) is a **subgroup** of \( G \).

As a consequence of proposition 863, the image of a group homomorphism is a subgroup of its range.

For an arbitrary subset \( A \) of \( G \), we denote the generated submodel by \( \langle A \rangle \). In addition to the elements of \( A \), \( \langle A \rangle \) contains their products and inverses, the products of their products and inverses, etc...

The **free group** builds a group out of a plain set; furthermore, as a consequence of ?? ([UNDEFINED]), every group is a quotient of a free group. Compare this to free semimodules and polynomial semirings.

(e) The **trivial group** is the trivial pointed set \( \{e\} \).

(f) We extend monoid exponentiation to all integers by setting

\[
x^{-n} := (x^n)^{-1}.
\]

This operation behaves well as shown in proposition 457 (e).

(g) The category of \( \mathcal{U} \)-small models of groups \( \mathcal{U} \text{-Grp} \) is concrete over \( \mathcal{U} \text{-Mon} \).

By proposition 457 (d), \( \mathcal{U} \text{-Grp} \) is also a concrete category over \( \mathcal{U} \text{-Inv} \).

The unique up to an isomorphism zero object in this category is the trivial group \( \{e\} \).

The **zero morphism** from \( G \) to \( H \) is

\[
0_{G,H} : G \to H
\]

\[
0_{G,H}(x) := e_H.
\]

We will define the free group functor in section 9.3 (Free groups). Then from proposition 917 it will follow that monomorphisms are precisely the injective homomorphisms, and that the categorical subobjects correspond to subgroups.

Unlike in the category \( \text{Mon} \) of monoids, in \( \text{Grp} \) every epimorphism is surjective.

We will prove this in proposition 508. Along with corollary 466, this shows that the categorical quotient objects correspond to quotient groups, which we will define shortly.

To avoid circularity, in this section, we will avoid using that monomorphisms are injective and epimorphism are surjective.
(h) The **kernel** of a group homomorphism \( \varphi : G \to H \) is the subgroup 
\[
\ker \varphi := \varphi^{-1}(e_H) = \{ x \in G \mid \varphi(x) = e_H \}.
\]
This coincides with the notion of a categorical kernel defined in definition 454 (b). Similarly to example 1217 (a), \( \ker \varphi \) is the equalizer of \( \varphi \) and the zero morphism \( 0_{G,H} \). This equivalence holds much more generally, even for pointed sets, however speaking of kernels is only established when we have an appropriate notion of cokernels. As we will see in definition 455 (i), cokernels are very well-behaved for groups, but not in general. Some related problems are highlighted in [Gol10, ch. 8]. Proposition 457 (h) expresses the compatibility between group kernels and cokernels.

(i) Consider the group homomorphism \( \varphi : G \to H \). We will find its **cokernel**. This will highlight several very fundamental facts about groups, especially quotient groups. In practice, quotients are conveniently characterized by theorem 464 (Quotient group universal property).

Similarly to example 1217 (b), the cokernel is an *equivalence partition* of \( H \). The partitioning relation is different, however. An equivalence relation that is compatible with the operations of an algebraic structure is called a **congruence**. For the group \( H \), the equivalence relation \( \cong \) is a congruence if:

- It is compatible with the group operation: \( x \cong x' \) and \( y \cong y' \) imply \( xy \cong x'y' \).
- It is compatible with identities: \( e_H \cong x \) implies \( y \cong xy \) for all \( y \in H \). This easily follows from the first condition.
- It is compatible with inverses: \( x \cong x' \) implies \( x^{-1} \cong x'^{-1} \). This also follows from the first condition: \( x^{-1} \cong x^{-1} \) implies \( e \cong x^{-1}x' \), and thus \( x'^{-1} \cong x^{-1} \).

We need congruences since we are working with groups and group homomorphisms rather than sets and functions. We define \( \cong \) to be the smallest congruence relation containing 
\[
\{(s(g), e_H) \mid g \in G\}.
\]
Denote the partition \( H/\cong \) by \( Q \). Define a group operation on \( Q \) as \( [x] \cdot [y] = [xy] \).

- This operation is well-defined since group congruences are compatible with the group operation. We are thus free to denote it via juxtaposition.
- The coset \( [e_H] \) is the identity of \( Q \) since congruences are compatible with identities.
- The coset \( [x]^{-1} \) is the inverse of \( [x] \) since congruences are compatible with inverses.

Therefore, \( Q \) is a group and \( \pi(x) := [x] \) is a group homomorphism. The pair \((Q, \pi)\) is thus a categorical cokernel of \( \varphi \) by the same argument as in example 1217 (b).

Denote the identity \( [e_H] \) by \( N \). It is a subgroup of \( H \):

- It contains the identity \( e_H \).
• It is closed under the group operation. Indeed, if \([x] = [y] = N\), then
  \[ [xy] = [x][y] = NN = N. \]

• It is closed under the group inverse. Indeed, \([x^{-1}] = N\) for every \(x \in H\). If \([x] = N\), then \([x^{-1}]N = N\), and hence \([x^{-1}] = N\).

• It possesses one additional important property. If \([x] = N\), then not only \(x \in N\), but also \(y^{-1}xy \in N\) for every \(y \in H\). This holds because
  \[ [y^{-1}xy] = [y^{-1}][x][y] = [y^{-1}][y] = [y]^{-1}[y] = N. \]

This last property distinguishes \(N\) from the image of \(\varphi\). A subgroup satisfying this property is called a normal subgroup. See ?? (UNDEFINED) for equivalent conditions. If the image \(\text{img } \varphi\) is a normal subgroup of \(H\), \(\varphi\) is a normal epimorphism in the sense of definition 454 (c).

Obviously \(\text{img } \varphi \subseteq N\). Since \(Q\) is a colimit, \(N\) must be the smallest normal subgroup containing \(\text{img } \varphi\).

It is more intriguing that \([x] = xN\) for every \(x \in H\). This can be shown as follows:

• Suppose first that \(y \in xN\), i.e. \(y = xn\) for some \(n \in N\). Then
  \[ y \in [y] = [xn] = [x]N = [x]. \]

  Generalizing on \(y\), we obtain that \(xN \subseteq [x]\).

• Conversely, let \(y \in [x]\). Obviously \(x = y(y^{-1}x)\). Then
  \[ [x^{-1}y] = [x^{-1}][y] = [x]^{-1}[y] = [x]^{-1}[x] = N. \]

  Hence, \(x^{-1}y \in N\) and \(y \in xN\). Generalizing on \(y\), we obtain that \([x] \subseteq xN\).

Therefore, all cosets in the quotient group \(Q = H/\sim\) are translations of the identity \(N\). In particular, in the notation of power set operations, it follows that

\[ xyN = xNyN. \]

Finally, given a normal subgroup \(N\) of an arbitrary group \(G\), we can define the quotient group \(G/N\) as the cokernel of the inclusion \(\iota : N \to G\). That is, \(G/N\) consists of the cosets \(xN\) for \(x \in G\) with the group operation \(xNyN = xyN\).

(j) If the only proper normal subgroup of \(G\) is the trivial subgroup \(\{e_G\}\), we say that \(G\) is a simple group.

The trivial group itself is not simple, because it has no proper subgroups.

**Example 456.** The power set magma \(\text{pow}(G)\) of a group \(G\) is a monoid, but it is not a group unless \(G\) is trivial.
Proposition 457. Every group $G$ has the following basic properties:

(a) The (binary) group operation is cancellative.

(b) The identity $e$ is its own inverse.

(c) $(xy)^{-1} = y^{-1}x^{-1}$.

(d) $x = (x^{-1})^{-1}$

(e) For any positive integer $n$, $(x^n)^{-1} = (x^{-1})^n$

(f) The map $x \mapsto x^{-1}$ is a group isomorphism.

(g) The kernel of a group homomorphism $\varphi : G \to H$ is trivial if and only if $\varphi$ is an embedding (injective homomorphism).

(h) For a quotient group $G/N$ with canonical projection $\pi(x) := xN$, the kernel of $\pi$ is $N$.

(i) The kernel of a group homomorphism is a normal subgroup.

Proof.

Proof of 457 (a). If $x = y$, obviously $xz = yz$ and $zx = zy$. Now if $xz = yz$, we have

$$x = x(zz^{-1}) = (xz)z^{-1} = (yz)z^{-1} = y(zz^{-1}) = y.$$ 

The case $zx = zy$ is analogous.

Proof of 457 (b). $ee = e$.

Proof of 457 (c).

$$(xy)(y^{-1}x^{-1}) = x(yy^{-1})x^{-1} = e = y^{-1}(x^{-1}x)y = (y^{-1}x^{-1})(xy).$$

Proof of 457 (d).

$$(x^{-1})^{-1} = xx^{-1}(x^{-1})^{-1} = x.$$

Proof of 457 (e). Using proposition 457 (d),

$$x^{-n} = (x^n)^{-1} = x^{-1}...x^{-1} = (x^{-1})^n.$$ 

Proof of 457 (f). Trivial.

Proof of 457 (g).

Proof of sufficiency. Suppose that $\ker \varphi = \{e_H\}$ and $\varphi(x) = \varphi(y)$. Then

$$e_H = \varphi(x)\varphi(y)^{-1} = \varphi(xy^{-1}).$$

Thus, $xy^{-1} \in \ker \varphi$, and hence $x = y$.

Therefore, $\varphi$ is injective.
Proof of necessity. Suppose that $\varphi$ is injective. Since $\varphi(e_G) = e_H$, $\varphi(x) = e_H$ implies that $x = e_G$.

Proof of 457 (h). Trivial.

Proof of 457 (i). For a homomorphism $\varphi : G \to H$, if $x \in \ker \varphi$, then
\[ \varphi(y^{-1}xy) = \varphi(y)^{-1} \varphi(x) \varphi(y) = \varphi(y)^{-1} \varphi(y) = e_H, \]
and thus $y^{-1}xy \in \ker \varphi$.

Proposition 458. A function between groups is a group homomorphism if and only if it satisfies (155).

Proof.

Proof of sufficiency. (155) is required to hold by definition.

Proof of necessity. Let the function $\varphi : G \to H$ satisfy (155). Then it preserves identities, i.e. is a pointed set homomorphism. Indeed, we have
\[ e_H \varphi(e_G) = \varphi(e_G) = \varphi(e_G e_G) = \varphi(e_G) \varphi(e_G). \]
By proposition 457 (a), $\varphi$ is cancellative, and hence $e_H = \varphi(e_G)$.
Inverses are preserved (i.e. (169) holds) because
\[ \varphi(x^{-1}) = \varphi(x^{-1})e_H = \varphi(x^{-1}) \varphi(x) \varphi(x)^{-1} = \varphi(x^{-1}x) \varphi(x)^{-1} = e_H \varphi(x)^{-1} = \varphi(x)^{-1}. \]
Therefore, $\varphi$ is indeed a group homomorphism.

Lemma 459. For each element $x$ of a group $G$, consider the function $\varphi_x := x \text{id}_G$, i.e.
\[
\varphi_x : G \to G \\
\varphi_x(y) := x \cdot y.
\]
This is a bijective function (but not necessarily a group isomorphism).

Proof.

Proof of injectivity. If $y, y' \in G$ and $\varphi_x(y) = \varphi_x(y')$, we have
\[ xy = \varphi_x(y) = \varphi_x(y') = xy'. \]
By proposition 457 (a), $y = y'$. Therefore, $\varphi_x$ is injective.

Proof of surjectivity. If $z \in G$, then $z = x(x^{-1}z)$. Therefore, $z = \varphi_x(x^{-1}z)$, and thus every member of $G$ has a preimage. Thus, $\varphi_x$ is surjective.

Proposition 460. The set of all invertible elements of a monoid is a group.

Proof. Fix a monoid $M$.

- $e_M$ is invertible.
• If \( x \) and \( y \) are invertible, then \( xy \) is invertible with inverse \( y^{-1}x^{-1} \).

• If \( x \) is invertible with inverse \( x^{-1} \), then \( x^{-1} \) is invertible with inverse \( x \).

Therefore, the set of invertible elements is a submonoid and is closed under inverses. \( \square \)

**Definition 461.** Let \( H \subseteq G \) be a subgroup of \( G \). Even if \( H \) is not normal, we can define the **left** and **right cosets**

\[
xH := \{xh \mid h \in H\} \quad Hx := \{hx \mid h \in H\}.
\]

The **index** \([G : H]\) of \( H \) in \( G \) is **cardinality** of the family of all left cosets.

The discussion in **definition 455 (i)** can be generalized to show that \( \{xH \mid x \in G\} \) is a **partition** of \( G \) into **equinumerous** sets. If \( H \) is not normal, this partition is not induced by a congruence, and we cannot form a quotient group using a non-normal subgroup. Nonetheless, left and right cosets still turn out useful.

**Proposition 462.** The family of all **left cosets** of a subgroup is **equinumerous** to the family of all right cosets.

**Proof.** Fix a subgroup \( H \) of \( G \), and consider the function \( xH \mapsto Hx \) taking left cosets to right cosets.

It is well-defined because, if \( xH = x'H \), then there exists \( h \), such that \( x = x'h \), and thus

\[
Hx' = Hh^{-1}x = Hx.
\]

It is injective by the same converse argument, and it is surjective by definition. Therefore, it is bijective. \( \square \)

**Proposition 463.** For a subgroup \( N \) of \( G \), the following conditions are equivalent:

(a) For every element \( x \) of \( G \), we have the set equality

\[
x^{-1}Nx = N.
\]

This is the definition of a normal subgroup obtained in **definition 455 (i)**.

(b) The partitions induced by the **left and right cosets** of \( N \) coincide.

(c) \( N \) is the kernel of some group homomorphism.

In particular, kernels are always normal subgroups.

**Proof.** This is the group-theoretic analog to **proposition 960**.

**Proof that 463 (a) implies 463 (b).** For any \( x \in G \)

\[
Nx = (xNx^{-1})x = xN(x^{-1}x) = xN,
\]

thus every left coset is a right coset and vice versa.
**Proof that 463 (b) implies 463 (c).** We can take the canonical projection \( \pi(x) := xN \) as the homomorphism. By proposition 457 (h), \( \ker \pi = N \).

**Proof that 463 (c) implies 463 (a).** Let \( \varphi : G \to H \) be a group homomorphism and fix any \( x \in G \). Denote \( N := \ker(f) \). By proposition 457 (i), it is a normal subgroup in the sense of definition 455 (i), i.e. it satisfies (170).

**Theorem 464** (Quotient group universal property). For every group \( G \) and normal subgroup \( N \), the quotient group \( G/N \) has the following universal mapping property:

Every group homomorphism \( \varphi : G \to H \) satisfying \( N \subseteq \ker \varphi \) uniquely factors through \( G/N \). That is, there exists a unique homomorphism \( \tilde{\varphi} : G/N \to H \), such that the following diagram commutes:

\[
\begin{array}{ccc}
G & \xrightarrow{\varphi} & H \\
\downarrow{\pi} & & \\
G/N & \xrightarrow{\tilde{\varphi}} & \\
\end{array}
\]

(171)

In the case where \( N = \ker \varphi \), \( \tilde{\varphi} \) is an embedding.

This extends to theorem 606 (Quotient module universal property) and theorem 649 (Quotient algebra universal property).

**Proof.** We want

\[ \tilde{\varphi}(\pi(x)) = \tilde{\varphi}(xN) = \varphi(x). \]

This suggests the definition

\[ \tilde{\varphi}(xN) := \varphi(x). \]

The homomorphism \( \tilde{\varphi} \) is well-defined because, if \( xN = x'N \), since \( N \subseteq \ker \varphi \), we have

\[ \varphi(x) = \varphi(x)e_N = \varphi(x)\varphi(N) = \varphi(xN) = \varphi(x'N) = \cdots = \varphi(y). \]

If \( N = \ker \varphi \), the kernel of \( \tilde{\varphi} \) is trivial. By proposition 457 (g), it is an injective function.

**Corollary 465.** Every group homomorphism \( \varphi : G \to H \) induces an isomorphism

\[ G/\ker \varphi \cong \text{img} \varphi. \]

**Proof.** Directly follows from theorem 464 (Quotient group universal property) by restricting the range of \( \tilde{\varphi} \) to its image.

**Corollary 466.** Every surjective group homomorphism is a normal epimorphism.

**Proof.** Fix a group homomorphism \( \varphi : G \to H \). By corollary 465, \( G/\ker \varphi \cong \text{img} \varphi = H \). Thus, \( H \) is a cokernel of the canonical inclusion \( \imath : \ker \varphi \to G \).

228
Theorem 467 (Quotient subgroup lattice theorem). Given a normal subgroup $N$ of $G$, the function $H \mapsto H/N$ is a lattice homomorphism between the lattice of subgroups of $G$ containing $N$ and the lattice of subgroups of the quotient $G/N$.

This extends to theorem 607 (Quotient submodule lattice theorem) and theorem 651 (Quotient ideal lattice theorem).

Proof.

Proof of injectivity. Let $H_1/N = H_2/N$. Both $H_1/N$ and $H_2/N$ consist of the same cosets, hence

$$H_1 = \bigcup (H_1/N) = \bigcup (H_2/N) = H_2.$$ 

Therefore, the map $H \mapsto H/N$ is injective.

Proof of surjectivity. Fix a subgroup $M$ of $G/N$ and define

$$H := \{ x \in G \mid xN \in M \}.$$ 

Then clearly $H/N = M$. Therefore, the map $H \mapsto H/N$ is surjective.

Proof of lattice compatibility. The join $\langle K \cup H \rangle$ of two subgroups of $G$ containing $N$ must satisfy the equality

$$\langle K \cup H \rangle/N = \langle (K/N) \cup (H/N) \rangle.$$ 

Verifying this amounts to noting that $\langle K \cup H \rangle$ is obtained by adding the products and inverses of any elements of $G$ not in $K \cup H$. Since the projection map $\pi : G \to G/N$ is a homomorphism, the coset $xyN$ of the product of $x, y \in \langle K \cup H \rangle$ is the product $(xN)(yN)$ of
the cosets $xN$ and $yN$, and analogously for inverses. Hence, adding an element $x \in G$ to 
$K \cup H$ and then taking all cosets is the same as adding the coset $xH$ to $(K/N) \cup (H/N)$. 
Therefore, (172) holds, and thus $H \mapsto H/N$ preserves joins in the lattice of subgroups. 
The other verifications are simpler. For meets, we have 

$$(K \cap H)/N = \{xN \mid x \in K \cap H\} = \{xN \mid x \in K\} \cap \{xN \cap x \in H\} = (K/N) \cap (H/N).$$

Finally, it remains to show that $H \mapsto H/N$ preserves the top and bottom elements. This 
is trivial since $G/N$ contains all possible cosets of $N$ and is hence the top in the lattice of 
subgroups of $G/N$, and $N/N$ is the trivial group and hence the bottom. 
Therefore, $H \mapsto H/N$ is a lattice isomorphism. 

**Theorem 468** (Lagrange’s theorem for groups). *Let $H$ be a subgroup of $G$. We have the 
following equality*

$$\text{card}(G) = \text{card}(H) \cdot [G : H]. \quad (173)$$

*If $H$ is a normal subgroup, then $[G : H] = \text{card}(G/H)$ and*

$$\text{card}(G) = \text{card}(H) \cdot \text{card}(G/H). \quad (174)$$

*This demonstrates that there exists a bijective function between the direct product $H \times G/H$ 
and $H$, however this may not be a group homomorphism — see example 518.*

**Proof.** By definition 461, every coset of $G$ with respect to $H$ is equinumerous with $H$, and 
there is a total of $[G : H]$ cosets. Therefore, (173) holds. 

**Example 469.** Consider the group $\mathbb{Z}$ of integers with respect to addition. 
Let $2\mathbb{Z}$ be the subgroup of all even integers. Then both $\mathbb{Z}$ and $2\mathbb{Z}$ are countably infinite, 
but their quotient group $\mathbb{Z}/2\mathbb{Z}$ has two elements — the set $2\mathbb{Z}$ of all even integers and the 
set $2\mathbb{Z} + 1$ of all odd integers. Generalizations of this quotient group are discussed in proposition 516. 
**Theorem 468** (Lagrange’s theorem for groups) holds, but it gives no insight due to the 
absorbing properties of transfinite cardinal arithmetic described in proposition 1083. 

Now consider the groups $4\mathbb{Z} \subseteq 2\mathbb{Z} \subseteq \mathbb{Z}$. As a consequence of **theorem 468** (Lagrange’s 
theorem for groups), $3\mathbb{Z}$ is not a subgroup of $2\mathbb{Z}$, and so we consider powers of 2. 

Since $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$, the quotient $2\mathbb{Z}/4\mathbb{Z}$ must a subgroup of $\mathbb{Z}/4\mathbb{Z}$ as a consequence of **theorem 467** (Quotient subgroup lattice theorem). We may not know the structure 
of the quotient groups (although we do, see **proposition 516**), but we know how $4\mathbb{Z}$, $2\mathbb{Z}$ and 
$\mathbb{Z}$ relate to each other, and we are able to determine how the quotient groups relate to each 
other.
9.3. Free groups

**Definition 470.** We associate with every set $A$ its **free monoid** $F(A) := (A^*, \cdot)$, where $A^*$ is the *Kleene star* and $\cdot$ is **concatenation**.

Denote by $\iota_A : A \to F(A)$ the canonical inclusion function, which sends every member of $A$ into the corresponding single-symbol word in the **free monoid** $F(A)$.

**Proof.** Concatenation is clearly associative and the empty word $\varepsilon$ is an identity under concatenation.

**Theorem 471 (Free monoid universal property).** We associate with every set $A$ its **Kleene star** $A^*$. Denote by $\iota_A : A \to F(A)$ the canonical inclusion function, which sends every member of $A$ into the corresponding single-symbol word in $A^*$.

The Kleene star $A^*$ with concatenation is the unique up to an isomorphism monoid that satisfies the following universal mapping property:

For every monoid $M$ and every function $f : A \to M$, there exists a unique monoid homomorphism $\bar{f} : A^* \to M$ such that the following diagram commutes:

$$
\begin{array}{ccc}
A & \xrightarrow{f} & M \\
\downarrow{\iota_A} & & \uparrow{\bar{f}} \\
F^* & & \\
\end{array}
$$

(175)

*Via remark 1198, $(-)^*$ becomes left adjoint to the forgetful functor

$U : Mon \to Set$.*

**Proof.** For every function $f : A \to M$, define the monoid homomorphism

$$
\bar{f} : A^* \to M,
$$

$$
\bar{f}(x_1x_2...x_n) := f(x_1) \cdot f(x_2) \cdot ... \cdot f(x_n)
$$

obtained by applying the monoid operation $\cdot$ recursively to the pointwise image

$$
f(x_1)f(x_2)...f(x_n)
$$

of the word

$$
x_1x_2...x_n.
$$

The homomorphism $\bar{f}$ is uniquely determined by the action of $f$ on single-symbol words.

**Definition 472.** Fix an arbitrary set $A$ and an **infix binary relation** $\rightarrow$ on $A$. We call the operation $\rightarrow$ a **reduction relation**, and the pair $(A, \rightarrow)$ an **abstract reduction system**.

In the case where $A$ is the Kleene star of some set, we also call $(A, \rightarrow)$ a **string rewriting system**.
We also introduce the following auxiliary relations:

- \( \cdot \) is the \( n \)-th iterated composition of \( \cdot \), where \( n \) is a nonnegative integer.
- \( + \) is the transitive closure of \( \cdot \).
- \( * \) is the reflexive closure of \( \cdot \).
- \( \leftrightarrow \) is the symmetric closure of \( \cdot \).
- \( + \leftrightarrow \) is the transitive closure of \( + \leftrightarrow \).
- \( * \leftrightarrow \) is the reflexive closure of \( + \leftrightarrow \), the smallest equivalence relation containing \( \cdot \).

In analogy with definition 1340 (c), if \( x \leftrightarrow y \), we say that \( y \) is a descendant of \( x \) and that \( x \) is an ancestor of \( y \). If \( x \rightarrow y \), we say that \( y \) is an immediate descendant, or a successor, especially when considering \( (A, \rightarrow) \) as a quiver.

We say that an element is reducible if it has an immediate descendant and irreducible otherwise.

We say that \( x \) and \( y \) are equivalent if \( x \leftrightarrow y \). This is a stronger condition than \( x \rightarrow y \) and \( y \rightarrow x \). Indeed, the latter corresponds to weak connectedness of quivers and equivalence corresponds to strong connectedness.

Given an equivalence relation \( \cong \), if \( x \cong y \) and \( y \) is irreducible, we say that \( y \) is a normal form of \( x \) modulo \( \cong \), and that \( x \) is normalizable modulo \( \cong \).

Without further context, we assume that \( \cong \) is the equivalence relation \( \leftrightarrow \).

If \( x \) has a unique normal form, we denote it by \( \text{nf} \).

**Definition 473.** Let \((A, \rightarrow)\) be an abstract reduction system. We will list several conditions ensuring existence and uniqueness of normal forms.

(a) We call the system confluent if, whenever \( x \) and \( y \) are descendants of \( w \), they have a common descendant \( z \). That is,

\[
\begin{array}{c}
\text{w} \\
\ast \quad \ast \\
\text{x} \quad \text{y} \\
\ast \ast \\
\text{z} 
\end{array}
\]

If there always exists an immediate descendant \( z \), we say that the system is locally confluent.

(b) We call the system noetherian or strongly normalizable if there exists no infinitely ascending sequence

\[
x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots.
\]

This is dual to well-foundedness.
Finally, we call the system **convergent** if it is both confluent and noetherian.

By proposition 474, this is equivalent to the system being **locally** confluent and noetherian.

**Proposition 474.** A noetherian abstract reduction system is confluent if and only if it is locally confluent.

**Proposition 475.** In a convergent abstract reduction system, every element has a unique normal form.

**Proof.** Let \((A, \to)\) be a convergent abstract reduction system.

**Proof of existence.** Suppose that \(x\) has no normal form. Hence, it is reducible and has an immediate descendant \(x_1\). Proceeding by natural number recursion, we can build a sequence

\[ x \rightarrow x_1 \rightarrow x_2 \rightarrow \cdots. \]

Such a sequence cannot exist, however, since the system is noetherian.

**Proof of uniqueness.** Suppose that both \(y\) and \(z\) are normal forms of \(w\).

Since the system is confluent, there exists an element \(z\) that is a descendant of both \(x\) and \(y\). But \(x\) and \(y\) are irreducible, therefore \(x = y\).

**Definition 476.** Let \(A\) be an arbitrary set. We will now construct the **free group** \(F(A)\) of \(A\).

Let \(\cdot\) be an arbitrary set not in \(A\). Consider the disjoint union \(U := A \times \{+, -\}\), whose members we will denote by \(a^+\) and \(a^-\).

On the Kleene star \(U\), define the relation \(\rightarrow\) to hold for \(x \rightarrow y\) if there exists \(a \in A\) and words \(p\) and \(s\) such that \(y = ps\) and either \(x = pa^+a^-s\) or \(x = pa^-a^+s\).

Then \((U, \rightarrow)\) is a convergent abstract reduction system. The **free group** \(F(A)\) is the set of irreducible elements in \(U\), with the operation \(v \star w := \text{nf}(vw)\) that gives the normal form of the concatenated string \(vw\).

The identity is the empty word and the inverse \(v^{-1}\) of \(v\) can be characterized recursively as

\[
v^{-1} := \begin{cases} 
\varepsilon & v = \varepsilon \\
\v^{-1}a^- & v = ua^+ \\
\v^{-1}a^+ & v = ua^- 
\end{cases}
\]

The canonical inclusion is

\[
\iota_A : A \rightarrow F(A) \\
\iota_A(a) := a^+.
\]

Compare this definition to free abelian groups defined in definition 551.

**Proof of correctness.** We need to prove that the system \((U, \rightarrow)\) is convergent. Then from proposition 475 it will follow that every word has a unique normal form.
**Proof of local confluence.** Fix two reductions \( w \to x \) and \( w \to y \).

If \( x = y \), define \( z := x \). Otherwise, let \( p_x \) and \( s_x \) be words such that \( w = p_x a^+ a^- s_x \) and \( x = p_x s_x \) (the case where \( w = p_x a^- a^+ s_x \) is analogous). Similarly, let \( p_y \) and \( s_y \) be words such that \( w = p_y b^+ b^- s_y \) and \( y = p_y s_y \).

Without loss of generality, suppose that \( p_x \) is a subword of \( p_y \). Then there exists a word \( v \) such that

\[
w = p_x a^+ a^- v b^+ b^- s_y.
\]

Then both \( x \) and \( y \) are reducible to \( z := p_x v s_y \).

This shows that the rewriting system is locally confluent.

**Proof of noetherianity.** All words are finite, and a reduction always removes two symbols. Hence, there cannot exist an infinite reduction path. \( \square \)

**Theorem 477** (Free group universal property). The free group \( F(A) \) is the unique up to an isomorphism group that satisfies the following universal mapping property:

\[
\begin{array}{ccc}
A & \xrightarrow{f} & G \\
\downarrow{\iota_A} & \searrow{\tilde{f}} & \\
F(A) & \quad & \\
\end{array}
\]

(177)

Via remark 1198, \( F \) becomes left adjoint to the forgetful functor

\[
U : \text{Grp} \to \text{Set}.
\]

**Proof.** The free group operation is more complicated than the free monoid operation, however the proof of the universal mapping property is identical to the property in theorem 471 (Free monoid universal property). \( \square \)

**Definition 478.** Let \( A \) be a plain set with free group \( F(A) \) and let \( R \subseteq F(A) \) be a set of words in \( F(A) \). Denote by \( N(R) \) the normal subgroup generated by \( R \).

Then we define the group with generators \( A \) and relators \( R \) as

\[
\langle A \mid R \rangle := F(A)/N(R).
\]

(178)

If \( R = \emptyset \), there are no relators, and we use the following notation for the free group:

\[
\langle A \rangle := F(A)
\]

(179)

Note that we use similar notation compared to generated subgroups defined in definition 455 (d). The former defines a new group operation from scratch, while the latter uses an existing group operation and is restricted by this operation.
If, for a given group $G$ and subsets $A$ and $R$ of $G$, we have

$$G \cong \langle A \mid R \rangle,$$

we call the pair $(A, R)$ a **presentation** of $G$.

We say that $G$ is **finitely generated** if both $A$ and $R$ are finite sets; if only $R$ is finite, we call $G$ **finitely presented**.

Compare this to module presentations and algebra presentations.

**Proposition 479.** Every group has at least one presentation.

Compare this to ?? ([UNDEFINED]) and ?? ([UNDEFINED]).

**Proof.** Fix an arbitrary group $G$ and let $A := U(G)$ be the underlying set.

By **Theorem 477** (Free group universal property), there exists a unique group homomorphism $\varphi : F(A) \to G$ such that

$$U(\varphi) \circ \iota_A = \text{id}_A.$$  

By **Proposition 457 (i)**, the kernel $\ker \varphi$ is a normal subgroup of $A$, hence by **Theorem 464** (Quotient group universal property),

$$G = \varphi(F(A)) \cong F(A)/\ker \varphi = \langle S \mid \ker \varphi \rangle.$$  

$\square$

**Definition 480.** The **free product** of a nonempty pairwise disjoint family of groups $\{X_k\}_{k \in \mathcal{K}}$ with presentations $\langle S_k \mid R_k \rangle, k \in \mathcal{K}$ is the group with presentation

$$\bigstar_{k \in \mathcal{K}} X_k := \left\langle \bigcup_{k \in \mathcal{K}} S_k \mid \bigcup_{k \in \mathcal{K}} R_k \right\rangle.$$  

If the constituent groups are not disjoint, we may instead use **disjoint unions** as

$$\bigstar_{k \in \mathcal{K}} X_k := \left\langle \left\{ (k, x) \mid k \in \mathcal{K} \text{ and } x \in S_k \right\} \bigcup \left\{ (k, x_1)(k, x_2) \ldots (k, x_n) : k \in \mathcal{K} \text{ and } x_1 x_2 \ldots x_n \in R_k \right\} \right\rangle.$$  

For every index $m \in \mathcal{K}$, we define the canonical embedding

$$t_m : X_m \to \bigstar_{k \in \mathcal{K}} X_k$$

$$t_m(x) := (m, x).$$

**Definition 481.** For a singleton alphabet $\{a\}$, we define the **infinite cyclic group** $C_\infty := \langle a \rangle$

and, for positive integers $n$, the **finite cyclic group of order** $n$ as

$$C_n := \langle a \mid a^n \rangle.$$  

We use the same notation independent of $a$ because all cyclic groups of the same order are obviously isomorphic.

Given an ambient group $G$ and some element $g \in G$, the **cyclic subgroup** of $g$ is the cyclic group isomorphic to the generated subgroup of $G$.

As shown in **Proposition 519**, cyclic groups are isomorphic to certain groups of integers, however it is still useful to have cyclic groups as a separate concept.
Definition 482. The order \( \text{ord}(G) \) of a group \( G \) is its cardinality.

The order \( \text{ord}(x) \) of a member \( x \) of a group is the smallest positive integer \( n \) such that \( x^n = e \), i.e. order of the cyclic subgroup generated by \( x \).

Proposition 483. Group orders have the following basic properties:

(a) For finite groups, the order of a group element divides the order of the group.

Proof of 483 (a). Follows from theorem 468 (Lagrange’s theorem for groups).

Proposition 484. The direct product \( C_m \times C_n \) of two cyclic groups is cyclic if and only if \( m \) and \( n \) are coprime.

Proof. The order of the element \((a, e)\) is \( m \) and the order of \((e, a)\) is \( n \). The order of \((a, a) = (a, e) \cdot (e, a)\) is the least common multiple of \( m \) and \( n \), which equals \( mn \) if and only if \( m \) and \( n \) are coprime.

Proposition 485. Groups of prime order are simple.

Proof. Let \( N \) be a proper normal subgroup of \( G \), where \( \text{ord}(G) \) is a prime number.

From theorem 468 (Lagrange’s theorem for groups) it follows that \( \text{ord}(N) \) divides \( p \). But \( p \) is prime, hence \( N \) is either the trivial group or the full group.

Therefore, \( G \) is a simple group.

Proposition 486.

(a) The categorical product of the family \( \{G_k\}_{k \in \mathcal{K}} \) in the category \( \text{Grp} \) of groups is their monoid direct product \( \prod_{k \in \mathcal{K}} G_k \).

(b) The categorical coproduct of the family \( \{G_k\}_{k \in \mathcal{K}} \) in the category \( \text{CMon} \) of commutative monoids is their group free product \( \ast_{k \in \mathcal{K}} G_k \).

Compare this to the commutative case discussed in proposition 451 (b).

Proof.

Proof of 486 (a). Follows from proposition 451 (a).

Proof of 486 (b). Let \( (A, \alpha) \) be a cocone for the discrete diagram \( \{G_k\}_{k \in \mathcal{K}} \). We want to define a group homomorphism \( l : \ast_{k \in \mathcal{K}} G_k \rightarrow A \) such that, for every \( m \in \mathcal{K} \),

\[
\alpha_m(x) = l_A(\iota_m(x)).
\]

This suggests the definition

\[
l_A((k_1, x_1)(k_2, x_2) ... (k_n, x_n)) := \alpha_{k_1}(x_1) \cdot \alpha_{k_2}(x_2) \cdot ... \cdot \alpha_{k_n}(x_n).
\]
9.4. Group actions

**Definition 487.** For every object $X$ in an arbitrary category $C$, the set $C(X)$ is a monoid with morphism composition as the monoid operation and $\text{id}_X$ as the monoid identity.

Outside of category theory, whenever the category $C$ is clear from the context, we call $C(X)$ the **endomorphism monoid** over $X$ and denote it by $\text{end}(X)$.

**Definition 488.** Let $M$ be a monoid and let $X$ be an object in some category $C$.

We will define monoid actions of $M$ on $X$, which we will sometimes call **left monoid actions**. There are also right monoid actions, which are only briefly mentioned in definition 488 (b).

A monoid action can be defined equivalently as:

(a) A homomorphism from $M$ to the endomorphism monoid $\text{end}(X)$.

(b) A functor from the delooping $B_M$ to $C$.

Right actions are contravariant functors.

(c) An indexed family $\{\Phi_m\}_{m \in M}$ of endomorphisms of $X$ such that

\[
\Phi_e = \text{id}_X, \quad \Phi_{mn} = \Phi_m \circ \Phi_n. \tag{MA2}
\]

This defines a function $\Phi : M \times A \to A$.

**Proof.**

**Proof that 488 (a) implies 488 (b).** Suppose that we have a monoid homomorphism $\Phi : M \to \text{end}(X)$. Define the functor

\[
F : B_M \to C
\]

\[
F(\cdot) := X
\]

\[
F(m) := \Phi(m).
\]

This is indeed a functor because (CF1) follows from (152) and (CF2) follows from (155).

**Proof that 488 (b) implies 488 (c).** Suppose that we have a functor $F : B_M \to C$. Let $X := F(\cdot)$ and define the $M$-indexed family

\[
\Phi_m : X \to X
\]

\[
\Phi_m := F(m).
\]

It satisfies the necessary axioms:

- **MA1** holds:

\[
\Phi_e = F(e)^{(CF1)} = \text{id}_A.
\]
• MA2 holds: for every pair \( m, n \in M \), we have
\[
\Phi_{mn} = F(mn) = F(m) \circ F(n) = \Phi_m \circ \Phi_n
\]

**Proof that 488 (c) implies 488 (a).** Suppose that we have an indexed family \( \{\Phi_m\}_{m \in M} \) of endomorphisms of \( A \) that satisfies the axioms for left action. Regard this indexed family as a function \( \Phi : M \to \text{end}(X) \).

Then \( \Phi \) is a monoid homomorphism because MA1 implies \( \Phi(e) = \text{id}_X \) and (MA2) implies
\[
\Phi(mn) = \Phi(m) \circ \Phi(n).
\]

**Proposition 489.** Every monoid acts on itself via the family of functions \( h \mapsto g \cdot h \) indexed by \( g \). These functions are not monoid homomorphisms in general.

Compare this result to theorem 496 (Cayley's theorem).

**Proof.** The family satisfies definition 488 (c):
- **MA1** follows from (165).
- **MA2** follows from associativity:
\[
[h \mapsto g_1 \cdot h] \circ [h \mapsto g_2 \cdot h] = [h \mapsto (g_1 \circ g_2) \cdot h].
\]

**Proposition 490.** The natural numbers \( \mathbb{N} \) (with zero) act on any monoid by exponentiation via the family of function \( g \mapsto g^n \) indexed by \( n \in \mathbb{N} \).

Compare this result to proposition 497.

**Proof.** This family satisfies definition 488 (c):
- **MA1** is obvious.
- **MA2** follows from proposition 444 (c).

**Definition 491.** For every object \( X \) in a groupoid \( G \), the set \( G(X) \) is a group with morphism composition as the group operation.

Similarly to endomorphism monoids, whenever the groupoid \( G \) is clear from the context, we call \( G(X) \) the automorphism group over \( X \) and denote it by \( \text{aut}(X) \).

**Definition 492.** We call the automorphism group of a set \( A \) the symmetric group\(^5\) on \( A \) and denote it by \( S(A) \). The group \( S(A) \) consists of bijective functions, which we call permutations.

Rather than for arbitrary sets, we often consider symmetric group
\[
S_n := S\{1, 2, \ldots, n\}.
\]

It is conventional to call \( S_n \) the “symmetric group on \( n \) letters”.

\(^5\)The term possibly comes from symmetric functions defined in definition 721
(a) It is common to write a permutation \(\sigma\) in \(S_n\) as
\[
\begin{pmatrix}
1 & \cdots & n \\
\sigma(1) & \cdots & \sigma(n)
\end{pmatrix}
\]

(b) If there exists a finite sequence \(k_1, \ldots, k_m\) of distinct numbers such that \(\sigma(k_m) = k_1\) and \(\sigma(k_{i+1}) = \sigma(k_i)\) for each \(i < m\), we say that the permutation is cyclic or a cycle of length \(m\). For brevity, we denote this cycle by \((k_1 \cdots k_m)\).

We call cycles of length 1 trivial cycles or loops and cycles of length 2 — transpositions. Every transposition is an involution, and thus a transposition is equal to its inverse permutation.

(c) If \((k_1 \cdots k_m)\) and \((s_1 \cdots s_l)\) are two cycles and if the sets \(\{k_1, \ldots, k_m\}\) and \(\{s_1, \ldots, s_l\}\) are disjoint, we say that the cycles themselves are disjoint.

Two disjoint cycles commute. That is,
\[
(k_1 \cdots k_m) \circ (s_1 \cdots s_l) = (s_1 \cdots s_l) \circ (k_1 \cdots k_m).
\]

**Proposition 493.** Every nontrivial cycle \((k_1 \cdots k_m)\) can be decomposed into the product of transpositions
\[
(k_1 \cdots k_m) = (k_1 \ k_m) \circ (k_1 \ k_{m-1}) \circ \cdots \circ (k_1 \ k_2).
\]

**Proof.** Trivial. \(\square\)

**Proposition 494.** The symmetric group \(S_n\) has \(n!\) elements.

**Proof.** We use induction on \(n\). The case \(n = 1\) is trivial. Suppose that \(S_{n-1}\) has \((n - 1)!\) elements. Then \(S_n\) is obtained by permuting \(n\) with each element of \(S_{n-1}\). That is,
\[
S_n = \{(k \ n) \circ \sigma \mid \sigma \in S_{n-1} \text{ and } 1 \leq k \leq n\}.
\]

It follows that
\[
\text{card}(S_n) = n \cdot \text{card}(S_{n-1}) = n(n-1)! = n!.
\]

**Definition 495.** Let \(G\) be a group and let \(X\) be an object in some concrete category \(\mathbf{C}\).

We will define group actions as a special case of monoid actions, with the same remarks regarding left and right group actions.

A group action can be defined equivalently as:

(a) A homomorphism from \(G\) to the automorphism group \(\text{aut}(X)\).

Right actions are homomorphisms from the opposite group \(G^{\text{op}}\) to \(\text{aut}(X)\).

(b) A functor from the delooping \(\mathbf{B}_G\) to \(\mathbf{C}\).

Right actions are contravariant functors.
(c) An indexed family \( \{\Phi_x\}_{x \in G} \) of isomorphisms of \( X \) such that, for every pair \( g, h \in G \),

\[
\Phi_{gh} = \Phi_g \circ \Phi_h.
\] (GA)

This defines a function \( \Phi : G \times A \to A \).

**Proof.** The proof of equivalence is simple; it is similar to definition 488. \( \square \)

**Theorem 496** (Cayley’s theorem). Every group acts on itself via the family of functions \( y \mapsto x \cdot y \) indexed by \( x \). These functions are not group homomorphisms in general.

Compare this result to **proposition 489**.

**Proof.** Follows directly from proposition 489 and lemma 459. \( \square \)

**Proposition 497.** The integers \( \mathbb{Z} \) act on any group by exponentiation via the family of function \( g \mapsto g^n \) indexed by \( n \in \mathbb{Z} \).

Compare this result to **proposition 490**.

**Proof.** Follows from proposition 490. \( \square \)

**Definition 498.** The orbit of \( x \) under the group action \( \Phi : G \to \text{end}(X) \) is the set

\[
\{\Phi_g(x) \mid g \in G\}.
\]

This is the set of all members of \( X \) “reachable” from \( x \) via the action.

The relation \( g \sim h \) on \( G \), defined to hold if \( g \) and \( h \) have the same orbit, is an equivalence relation. The quotient set \( G/\sim \) is a partition of \( G \) into sets called **orbits**.

**Example 499.** Consider the action of the additive group of \( \mathbb{R} \) on \( \mathbb{R}^2 \) given by the rotation matrices

\[
\Phi_r := \begin{pmatrix}
\cos(r) & \sin(r) \\
-\sin(r) & \cos(r)
\end{pmatrix}
\]

Fix a nonzero vector \( (x, y)^T \) in \( \mathbb{R}^2 \) with norm \( l \). Since rotation matrices are orthogonal, they preserve norms. Furthermore, given a vector of norm \( l \), with the angle \( r \) defined via (142), \( \Phi_r^{-1} \) sends the vector to \( (x, y)^T \).

The orbit of \( (x, y)^T \) is thus a **circle** at the origin with radius \( l \).

**Proposition 500.** Every group acts on itself via the **conjugation automorphisms** defined as

\[
\Phi_g : G \to G,
\]

\[
\Phi_g(h) := ghg^{-1}.
\]

**Proof.** Trivial. \( \square \)

**Definition 501.** Denote by \( \Phi : G \to \text{aut}(G) \) the conjugation action on the group \( G \).

The **inner automorphisms group** of \( G \) is

\[
\text{inn}(G) := \{\Phi_g \mid g \in G\}.
\]

The **outer automorphism group** is the quotient group

\[
\text{out}(G) := \text{aut}(G)/\text{inn}(G).
\]
Proof of correctness. We will show that \( \text{inn}(G) \) is a normal subgroup of \( \text{aut}(G) \).

Fix a member \( g \) of \( G \) and define the inner automorphism

\[
\varphi(h) := gh^{-1}g.
\]

Let \( \psi \) be an arbitrary automorphism of \( G \). Then

\[
[\varphi \circ \psi \circ \varphi^{-1}](h) = \varphi(\varphi^{-1}(h)g^{-1}) = \varphi(g)h\varphi(g^{-1}),
\]

and thus \( \varphi \circ \psi \circ \varphi^{-1} \) is again an inner automorphism. \( \square \)

**Proposition 502.** Let \( S_n \) be a symmetric group and let \( \sigma \) be a permutation in \( S_n \). Then there exists a finite sequence \( c_1, ..., c_m \) of nontrivial disjoint cycles such that

\[
\sigma = c_1 \circ \cdots \circ c_m.
\]

The case where \( \sigma \) is the identity permutation corresponds to \( m = 0 \).

We call this a cycle decomposition of \( \sigma \). Via proposition 493, this permutation can further be decomposed into a composition of individual transpositions. We call the latter a transposition decomposition of \( \sigma \).

Neither decomposition is unique. Nonetheless, existence is sufficient for most practical purposes, including permutation parity defined in definition 504.

**Proof.** Suppose that \( \sigma \) is not the identity.

Consider the group action

\[
\Phi(\sigma) : \langle \sigma \rangle \times \{1, ..., n\} \to \{1, ..., n\}
\]

of the generated subgroup \( \langle \sigma \rangle \).

The orbit of \( k \) under \( \Phi(\sigma) \) is the set

\[
O_{\sigma}(k) := \{\sigma^m(k) \mid m \in \mathbb{Z}\}
\]

of all numbers reachable from \( k \) via iterated application of \( \sigma \) or \( \sigma^{-1} \). The family

\[
\{O_{\sigma}(k) \mid k = 1, ..., n\}
\]

of orbits partitions \( 1, ..., n \) into disjoint subsets.

Each orbit \( O \) has a smallest element \( o \); and \( O_{\sigma}(o) = O \). This smallest element uniquely identifies a cycle

\[
(o \ \sigma(o) \ \sigma^2(o) \ \cdots \ \sigma^{c-1}(o)),
\]

where \( c \) is the cardinality of \( O_{\sigma}(o) \).

Ignoring the trivial cycles, we obtain a unique set of disjoint nontrivial cycles for every permutation \( \sigma \) in \( S_n \), which we call the cycle decomposition of \( \sigma \). Denote this set by \( C_{\sigma} \).

We must prove that \( C_{\sigma} \) is family of disjoint sets. Let \( O_{\sigma}(k_1) \) and \( O_{\sigma}(k_2) \) be two orbits.
Suppose that there exists a number \( k \in O_\sigma(k_2) \cap O_\sigma(k_1) \). Then there exist \( m_1, m_2 \leq n \) such that \( k = \sigma^{m_1}(k_1) \) and \( k = \sigma^{m_2}(k_2) \). Then

\[ k_1 = \sigma^{-m_1}(k) = \sigma^{-m_1}(\sigma^{m_2}(k_2)) = \sigma^{m_2-m_1}(k_2), \]

and thus \( O_\sigma(k_2) \subseteq O_\sigma(k_1) \). We can analogously obtain the converse inclusion.

Therefore, if two orbits have a nonempty intersection, they are equal.

Finally, we will use induction on \( m \) to show that these cycles give us \( \sigma \). The case \( m = 0 \) is trivial. For the inductive hypothesis, note that \( c_1^{-1} \sigma \) has as cycles \( c_2, \ldots, c_m \) because all the numbers in \( c_1 \) are fixed points of \( c_1^{-1} \sigma \). Thus, the inductive hypothesis holds for \( c_1^{-1} \sigma \).

From

\[ c_1^{-1} \sigma = c_2 \circ \cdots \circ c_m \]

it follows that

\[ \sigma = c_1 \circ \cdots \circ c_m. \]

\[ \square \]

**Lemma 503.** If a permutation \( \sigma \in S_n \) can be decomposed into transpositions as both

\[ \sigma = (k_1 \ k_2) \circ (k_3 \ k_4) \circ \cdots \circ (k_{2n-1} \ k_{2n}) \quad (180) \]

\( n \) transpositions

and

\[ \sigma = (l_1 \ l_2) \circ (l_3 \ l_4) \circ \cdots \circ (l_{2m-1} \ l_{2m}), \]

\( m \) transpositions

then \( n - m \) is an even number.

**Proof.** We will use induction on \( n \). First consider the base case \( n = 0 \). Then \( \sigma \) is the identity. Hence, every transposition in (181) should be present twice so that its action cancels out. Therefore, \( m \) is an even number.

Now suppose that the statement holds for \( n - 1 \). Add (compose on the right) the last transposition \( (k_{2n-1} \ k_{2n}) \) of (180) to both (180) and (181). The obtained permutations are obviously equal. Furthermore, since \( (k_{2n-1} \ k_{2n}) \) is its own inverse, we can just as well remove \( (k_{2n-1} \ k_{2n}) \) from (180) to obtain a decomposition into \( n - 1 \) (rather than \( n + 1 \)) transpositions.

By the inductive hypothesis, \( (n - 1) - (m + 1) = n - m - 2 \) is an even number. Therefore, \( n - m \) is also an even number. \[ \square \]

**Definition 504.** We say that a permutation \( \sigma \in S_n \) is **even** (resp. **odd** if a decomposition of \( \sigma \) into transpositions has an even (resp. odd) number of transpositions.

By *proposition 502*, such a decomposition exists. By *lemma 503*, all such decompositions yield the same parity even if they decompose into a differing number of transpositions.

We correspondingly define the **sign** of a permutation as

\[ \text{sgn} : S_n \rightarrow \mathbb{Z}, \]

\[ \text{sgn}(\sigma) := \begin{cases} 1, & \text{\sigma is even} \\ -1, & \text{\sigma is odd} \end{cases} \]

242
**Definition 505.** The alternating group $A_n$ on $n$ letters is the subgroup of all even permutation in the symmetric group $S_n$.

**Proposition 506.** The alternating group $A_n$ has $n!/2$ elements.

**Proof.** The proof is similar to that of proposition 494, but is a little different.

We use induction on $n$. The case $n = 1$ is trivial. Suppose that $A_{n-1}$ has $(n-1)!/2$ elements. Then

$$A_n = \{(k \ n) \circ \sigma \mid \sigma \in S_{n-1} \setminus A_{n-1} \text{ and } 1 \leq k \leq n\}.$$

We obtain $A_n$ by taking all the odd permutations in $S_{n-1}$ and composing them with one new transposition. It follows that

$$\text{card}(A_n) = n \cdot \text{card}(S_{n-1} \setminus A_{n-1}) = n \frac{(n-1)!}{2} = \frac{n!}{2}.$$

\[\square\]

**Example 507.** The symmetric group $S_3$ contains the following permutations:

$$S_3 := \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \right\}.$$

We can observe the following:

- The permutations $(1 \ 2 \ 3)$ and $(1 \ 3 \ 2)$ are inverses of each other and all other permutations are involutions.

- Every conjugation automorphism is unique. This can be verified explicitly. Therefore, the inner automorphism group $\text{inn}(S_3)$ is isomorphic to $S_3$.

- The alternating group $A_3$ consists of the identity and the odd-length cycles $(1, 2, 3)$ and $(1, 3, 2)$.

- When restricted to $A_3$, all conjugation automorphisms are trivial. This can be verified explicitly. Therefore, the inner automorphism group $\text{inn}(A_3)$ is trivial, and hence

$$\text{aut}(A_3) \cong \text{out}(A_3).$$

- The map $\sigma \mapsto \sigma^{-1}$, which fixes the identity and exchanges the two other permutations, is an automorphism of $A_3$. It is distinct from the identity, hence it is an outer automorphism.

This map is given by the restriction of the conjugation $\sigma \mapsto (1 \ 2)\sigma(1 \ 2)$ to $A_3$. It is an inner automorphism of $S_3$, but an outer automorphism of $A_3$.

**Proposition 508.** Every epimorphism in $\text{Grp}$ is surjective.
Proof. Let \( \varphi : G \to H \) be an epimorphism and suppose that it is not surjective. Let \( M \) be the smallest normal subgroup of \( H \) containing \( \text{img} \varphi \).

If \( M \) has index 2 in \( H \), consider the quotient map \( \pi : H \to H/M \) and the constant map \( c(h) := M \). Then
\[
\pi \circ \varphi = c \circ \varphi.
\]

Since \( \varphi \) is an epimorphism, we have \( \pi = c \). But we have deliberately taken \( \pi \) and \( c \) so that \( \pi \neq c \). The obtained contradiction shows that \( M \) must have an index greater than 2.

Let \( M, uM \) and \( vM \) be different cosets. Define \( \sigma : H \to H \) as the permutation on \( H \) that exchanges \( xu \) with \( xv \) for every \( x \in M \). Define the homomorphism
\[
\psi : H \to S(H)
\]
\[
\psi(h) := (x \mapsto hx),
\]
where \( S(H) \) is the symmetric group.

This is indeed a homomorphism by theorem 496 (Cayley’s theorem). By proposition 500, another homomorphism is
\[
\theta : H \to S(H)
\]
\[
\theta(h) := \sigma^{-1} \circ \psi(h) \circ \sigma.
\]

Since \( \sigma \) fixes the members of \( M \) in-place, we have \( \theta(h)|_M = \psi(h)|_M \). Since \( M \) contains the image of \( \varphi \), this implies
\[
\psi \circ \varphi = \theta \circ \varphi.
\]

Since \( \varphi \) is an epimorphism, we have \( \psi = \theta \). But we have deliberately constructed \( \psi \) and \( \theta \) such that \( \psi \neq \theta \). The obtained contradiction shows that \( \text{img} \varphi \) cannot be a strict subgroup of \( G \). Therefore, \( \varphi \) must be surjective. \( \square \)

**Definition 509.** Suppose that \( C \) is a concrete category and let \( X \) be an object of \( C \).

A **dynamical system** is a monoid action \( \Phi : T \times X \to X \). We call \( X \) the **phase space** of the system. In applications, we interpret the monoid \( T \) as **time** and consider it to be additive. We call \( \Phi \) the **evolution function** of the system.

(a) If \( T \) is either the additive monoid of the natural numbers or the additive group of the integers, we say that the dynamical system has **discrete time**.

Due to (MA2), \( \Phi_{n+1} = \Phi_n \circ \Phi_1 \) for any integer \( n \). Using natural number induction and proposition 457 (e), we can show that \( \Phi_n = \Phi_1^n \) for every integer \( n \).

Therefore, the entire evolution function of a discrete-time dynamical system is determined by a single endofunction \( \varphi : X \to X \). For this reason, we also refer to discrete-time dynamical systems as **iterated function systems**.

(b) If \( T \) is the additive monoid of real numbers, with or without an infinite element, we say that the system is a **semiflow**.

(c) If \( T \) is the additive group of real numbers, we say that the system is a **flow**.
We will call the system **discrete** if $T$ is the monoid of zero-based **natural numbers** and **continuous** if $T$ is the monoid of nonnegative **real numbers**, with or without an **infinite element**.

The monoid $T$ can theoretically be a **group**, in which case we consider **group actions**, however negative time is not as often needed in practice.

**Definition 510.** Fix a dynamical system with evolution function $\Phi : T \times X \to X$.

A **trajectory** in a starting at the **initial state** $x_0 \in X$ is an indexed family $\{x_t\}_{t \in T}$ obtained as

$$x_t := \Phi_t(x_0).$$

The condition **MA1** ensures that $\Phi_0(x_0) = x_0$, and **MA2** ensures that

$$x_{t+s} = \Phi_{t+s}(x_0) = \Phi_t(x_s).$$

For discrete dynamical systems, trajectories are sequences.
9.5. Abelian groups

Definition 511. A commutative group is usually called an abelian group. We denote by \( \text{Ab} \) the category of abelian groups.

By proposition 648, the abelian groups are precisely the rings over \( \mathbb{Z} \), and we have an isomorphism of categories \( \text{Ab} \cong \text{Mod}_\mathbb{Z} \).

Remark 512. General groups often arise as automorphism groups, which are, for the most part, non-commutative, while abelian groups are usually used as the main building block for rings and modules.

To make a further distinction, if the operation is denoted by \( \cdot \) or juxtaposition, we say that the group is a multiplicative group, and if the operation is denoted by \( + \), we say that the group is an additive group. This terminology usually, but not necessarily, coincides with the group (or, more generally, the magma) being commutative.

To make things explicit, a multiplicative magma is any magma as defined in definition 441. Compare this to additive magmas, where

(a) The magma operation is denoted by \( + \) and called addition.

(b) The magma exponentiation operation \( x^n \) is denoted by \( n \cdot x \) or juxtaposition and called multiplication. Thus, multiplication is not defined for two elements of the magma, but defined for a positive integer and an element of the magma. That is,

\[
\cdot : \cdot : \mathbb{N} \times R \to R \quad \text{with} \quad n \cdot x := \begin{cases} 
0_M, & n = 0, \text{initial condition if } M \text{ is a monoid} \\
x, & n = 1, \text{initial condition if } M \text{ is not a monoid} \\
n \cdot x + x, & n > 1 \\
-(n \cdot x), & n < 0, 
\end{cases} \tag{182}
\]

In the case of a commutative monoid, if multiplication is extended to two elements of the monoid, we instead talk about semirings.

(c) The identity is usually denoted by 0.

(d) If an inverse of \( x \) exists, it is denoted by \(-x\) rather than \( x^{-1} \).

Proposition 513. In an abelian group, the full automorphism group \( \text{aut}(G) \) is isomorphic to the outer automorphism group \( \text{out}(G) \).

Proof. If the group operation is commutative, then \( xyx^{-1} = yxx^{-1} = y \), which makes the conjugation action trivial. Thus, the inner automorphism group \( \text{int}(G) \) is trivial, and hence \( \text{aut}(G) \cong \text{out}(G) \).

Proposition 514. All subgroups of an abelian group are normal.

Proof. Let \( G \) be abelian and \( H \) be a subgroup of \( G \). Then \( xHx^{-1} = xx^{-1}H = H \) for any \( x \in G \) and thus \( H \) is normal.
Definition 515. Given a normal subgroup \( N \) of an abelian group \( G \), we say that two elements \( x \) and \( y \) of \( G \) are congruent modulo \( N \) and write \( x \equiv y \pmod{N} \) if \( x - y \in N \).

If \( N = \langle z \rangle \), this implies that \( x \equiv y \pmod{z} \) if and only if \( x - y \in \langle z \rangle \).

This concept also extends to ring ideals rather than normal subgroups, in which case \( \langle z \rangle \) is the ideal generated by \( z \) rather than the cyclic subgroup of \( z \).

Proposition 516. The integers \( \mathbb{Z} \) form an abelian group under addition. For every positive integer \( n \), we define the group
\[
\mathbb{Z}_n := \{0, 1, \ldots, n - 1\}
\]
with the operation
\[
x \oplus y := \text{rem}(x + y, n)
\]
so that
\[
x \oplus y \equiv x + y \pmod{n}.
\]

The group \( \mathbb{Z}_n \) is called the group of integers modulo \( n \). Compare this result with proposition 581.

Proof. We will prove that \( \mathbb{Z}_n \) is an abelian group.

Proof of associativity. Addition in \( \mathbb{Z}_n \) is associative since
\[
(x \oplus y) \oplus z = \text{rem}((x \oplus y) + z, n) = \\
= \text{rem}((x + y, n) + z, n) = \\
= \text{rem}(x + y - n \text{quot}(x + y, n) + z, n) = \\
= \text{rem}(x + y + z, n) = \\
= \ldots = \\
= x \oplus (y \oplus z).
\]

Proof of identity. The zero is the identity.

Proof of inverse. Fix \( x \in \mathbb{Z}_n \). If \( x = 0 \), its inverse is 0. If \( x > 0 \), its inverse is \( n - x \) since \( n - x \in \mathbb{Z}_n \) and
\[
x \oplus (n - x) = x + (n - x) - n = 0.
\]

Proof of commutativity. Follows from
\[
x \oplus y = \text{rem}(x + y, n) = \text{rem}(y + x, n) = y \oplus x.
\]

\( \square \)

Proposition 517. The group \( \mathbb{Z}_n \) of integers modulo \( n \) is isomorphic to the quotient of \( \mathbb{Z} \) by \( n\mathbb{Z} = \{nz \mid z \in \mathbb{Z}\} \). That is,
\[
\mathbb{Z}_n \cong \mathbb{Z}/n\mathbb{Z}.
\]
Proof. Define the function
\[
\varphi : \mathbb{Z}_n \to \mathbb{Z}/n\mathbb{Z}
\]
\[
\varphi(x) := x + n\mathbb{Z}.
\]

It is a homomorphism because
\[
\varphi(x \oplus y) = \varphi(\text{rem}(x + y, n)) = \\
= \varphi(x + y - n \text{quot}(x + y, n)) = \\
= x + y - n \text{quot}(x + y, n) + n\mathbb{Z} = \\
= x + y + n\mathbb{Z} = \\
= (x + n\mathbb{Z}) + (y + n\mathbb{Z}) = \\
= \varphi(x) + \varphi(y).
\]

Furthermore, this shows that \( \varphi \) is also an isomorphism. \( \square \)

**Example 518.** Theorem 468 (Lagrange's theorem for groups) and proposition 517 imply that, for any positive integer \( n \), \((nm, k) \mapsto nm + k\) is a bijection between \( n\mathbb{Z} \times \mathbb{Z}_n \) and \( \mathbb{Z} \). This bijection, however, is not necessarily a group isomorphism because (155) may not hold.

Consider the tuples \((nm_1, k_1)\) and \((nm_2, k_2)\) in \( n\mathbb{Z} \times \mathbb{Z}_n \). We have
\[
(nm_1, k_1) + (nm_2, k_2) = (nm_1 + nm_2, \text{rem}(k_1 + k_2, n)).
\]

Therefore, if \( k_1 + k_2 \geq n \),
\[
nm_1 + nm_2 + \text{rem}(k_1 + k_2, n) < (nm_1 + k_1) + (nm_2 + k_2).
\]

**Proposition 519.** The cyclic group \( C_n \) is isomorphic to the group \( \mathbb{Z}_n \) of integers modulo \( n \).

Proof. The homomorphism
\[
\varphi : \mathbb{Z}_n \to C_n
\]
\[
\varphi(k) := a^k,
\]
and the analogous homomorphism for the infinite group, are isomorphisms. \( \square \)

**Definition 520.** Let \( M \) be a commutative monoid. Define the equivalence relation \( \sim \) on tuples of members of \( M \) to hold for \((a, b) \sim (a', b')\) if there exists an element \( u \) of \( M \) such that
\[
a + b' + u = a' + b + u.
\]

Define addition on the equivalence partition \( G := (M \times M)/\sim \) componentwise as
\[
[(a, b)] \oplus [(c, d)] := [(a + c, b + d)]
\]
and fix a canonical embedding
\[
t_M : M \to G
\]
\[
t_M(m) := [(m, 0)].
\]

We call the obtained abelian group \( (G, \oplus) \) the Grothendieck completion of \( M \).
Proof of correctness.

Proof that $\sim$ is an equivalence relation.

Proof of reflexivity. \[(a, b) \sim (a, b)\] if and only if \[a + b + 0 = a + b + 0\]

Proof of symmetry. By commutativity, if \((a, b) \sim (a', b')\), then there exists \(u\) such that
\[a + b' + u = a' + b + u = a' + b + u = a + b' + u,\]

hence \((a', b') \sim (a, b)\).

Proof of transitivity. Suppose that \((a, b) \sim (a', b')\) and \((a', b') \sim (a'', b'')\). Thus, there exist elements \(u\) and \(b\) of \(M\) such that
\[a + b' + u = a' + b + u = a' + b + u = a + b' + u,\]

\[a' + b'' + v = a'' + b' + v.\]

Summing both sides, we obtain
\[(a + b' + u) + (a' + b'' + v) = (a' + b + u) + (a'' + b' + v)].\]

We reorder both sides to obtain
\[(a + b'') + (a' + b' + u + v) = (a'' + b) + (a' + b' + u + v),\]

which implies \((a, a'') \sim (b, b'').\)

Proof that \((G, \oplus)\) is an abelian group.

Proof that \(\oplus\) is well-defined. The addition operation on \(G\) does not depend on the representative of the equivalence class. Indeed, let \((a, b) \sim (a', b')\) and \((c, d) \sim (c', d')\). Then there exist \(u\) and \(b\) such that
\[a + b' + u = a' + b + u,\]
\[c + d' + v = c' + d + v.\]

When added combined, these give
\[(a + c) + (b' + d') + (u + v) = (a' + c') + (b + d') + (u + v),\]

which implies that
\[(a + c, b + d) \sim (a' + c', b' + d').\]

Proof of associativity. Associativity of multiplication in \(G\) is inherited from multiplication in \(M\).

Proof of identity. The equivalence class \([(0, 0)]\) is an identity in \(G\) and contains the pairs \((x, x)\) of identical elements.
**Proof of inverse.** For each member \((a, b) \in M \times M\), its inverse is \((b, a)\) because

\[
[(a, b)] \oplus [(b, a)] = [(a + b, b + c)],
\]

which, by commutativity, belongs to \([(0, 0)]\).

**Proof of commutativity.** Commutativity of the group operation \(\oplus\) is also inherited from the monoid operation \(+\).

\[\square\]

**Theorem 521** (Grothendieck monoid completion universal property). The Grothendieck completion \(\overline{M}\) of a commutative monoid \(M\) satisfies the following universal mapping property:

For every abelian group \(G\) and every monoid homomorphism \(\varphi : M \to G\), there exists a unique group homomorphism \(\overline{\varphi} : \overline{M} \to G\) such that the following diagram commutes:

\[
\begin{array}{ccc}
M & \overset{\varphi}{\longrightarrow} & G \\
\downarrow & & \downarrow \\
\overline{M} \quad & \overset{\overline{\varphi}}{\longrightarrow} & G
\end{array}
\]

Via remark 1198, \(\overline{\varphi}\) becomes left adjoint to the forgetful functor

\[U : \text{Ab} \to \text{CMon}.\]

Compare this result to theorem 587 (Grothendieck semiring completion universal property).

**Proof.** Let \(\varphi : M \to G\) be a monoid homomorphism into an abelian group \(G\). We want to define a homomorphism \(\overline{\varphi}\) such that

\[
\overline{\varphi}(t_M(a)) = \overline{\varphi}([(a, 0)]) = \varphi(a).
\]

Each equivalence class \(C\) in \(G\) has a unique member \(a\) such that \((a, 0) \in C\), hence the above condition is well-posed.

Fix pairs \((a, b)\) and \((a', b')\) from \(M \times M\). Suppose that \((a, b) \sim (a', b')\). Then there exists \(u \in M\) such that

\[a + b' + u = a' + b + u.\]

An additional restriction on \(\overline{\varphi}\) is then

\[\overline{\varphi}([(a, b)]) = \overline{\varphi}([(a', b')]).\]

We need to cancel out \(u\). This uniquely determines \(\overline{\varphi}\) as

\[\overline{\varphi}([(a, b)]) := \varphi(a) - \varphi(b).\]

\[\square\]
**Definition 522.** Let $G$ be an arbitrary group. We define the **commutator** of the elements $x$ and $y$ as

$$[x, y] := x y x^{-1} y^{-1}.$$ 

The **commutator subgroup** $[G, G]$ of $G$ is the subgroup generated by all the commutators in $G$.

**Theorem 523** (Group abelianization universal property). The commutator group $[G, G]$ of any group $G$ is normal and the quotient $G/[G, G]$ is an abelian group, which we call the **abelianization** of $G$, satisfies the following universal mapping property:

For every abelian group $H$, every group homomorphism $\phi : G \to H$ uniquely factors through $G/[G, G]$. That is, there exists a unique group homomorphism $\bar{\phi} : G/[G, G] \to H$ such that the following diagram commutes:

$$
\begin{array}{ccc}
G & \xrightarrow{\phi} & H \\
\downarrow{\pi_G} & & \downarrow{\bar{\phi}} \\
G/[G, G] & & 
\end{array}
$$

(184)

Via remark 1198, the abelianization functor becomes left adjoint to the forgetful functor $U : \text{Ab} \to \text{Grp}$.

This result extends to theorem 590 (Ring abelianization universal property).

**Proof.** Let $C := [G, G]$.

**Proof that $G/C$ is abelian.** Normality of $G/C$ easily follows from

$$a x y x^{-1} y^{-1} a^{-1} = (a x a^{-1})(a y a^{-1})(a x a^{-1})^{-1}(a y a^{-1})^{-1}.$$ 

Then for the cosets $a C$ and $b C$, we have

$$a C \cdot b C = a b C = a b (b^{-1} a^{-1} b a) C = b a C.$$ 

Therefore, the quotient group $G/C$ is abelian.

**Proof of universal mapping property.** Let $H$ be an abelian group and let $\phi : G \to H$ be a group homomorphism.

Observe that $\phi(C) = e_H$. Indeed, since $H$ is abelian, for $[x, y] = x y x^{-1} y^{-1} \in C$ we have

$$\phi([x, y]) = \phi(x) \phi(y) \phi(x^{-1}) \phi(y^{-1}) = \phi(x) \phi(x^{-1}) \phi(y) \phi(y^{-1}).$$

We want $\bar{\phi} : G/C \to H$ to satisfy

$$\bar{\phi}(\pi_G(x)) = \phi(x).$$
This suggests the definition

$$\varphi(xC) := \varphi(x).$$

It is well-defined because if $xC = yC$, we have

$$\varphi(x) = \varphi(x)e_H = \varphi(x)\varphi(C) = \varphi(xC) = \varphi(yC) = \ldots = \varphi(y).$$
10. Ring theory

As discussed in remark 512, commutative and non-commutative groups are quite different despite having similar definitions. Rings are extensions of abelian groups, which allow multiplication with more than members of \( \mathbb{Z} \).

For commutative rings, this second operation is often truly an extension of definition 441 (e) to arbitrary ring elements. For noncommutative ring, this second operation is usually given by function composition.

This section also describes modules, which are important both as generalizations of vector spaces and as a tool to study rings. Ring ideals are instances of submodules, for example.

In an attempt to encompass the natural numbers, lattice ideals and polynomials and matrices over tropical semirings, we have chosen to use semirings, semimodules and semiring ideals as fundamental notions.

Figure 20: Some important semirings

- semiring
- noetherian
- zerosumfree
- commutative
- entire
- ring
- distributive lattice
- integral domain
- division ring
- simple ring
- greatest common divisor domain
- unique factorization domain
- principal ideal domain
- Euclidean domain
- field
10.1. Semirings

We will start by defining semirings, and to do that we will first motivate distributivity.

**Proposition 524.** Fix an additive monoid \((R, +, \cdot)\), where \(+: R \times R \rightarrow R\) is the monoid operation and \(\cdot : \mathbb{N} \times R \rightarrow R\) is defined via (182).

We have the following property, which we call **distributivity** of \(\cdot\) over \(+\):

\[
    n \cdot (x + y) = n \cdot x + n \cdot y. \tag{185}
\]

**Proof.** We use induction on \(n\). The case \(n = 0\) is trivial. Suppose that (185) holds. Then

\[
    (n + 1) \cdot (x + y) \overset{(186)}{=} n \cdot (x + y) + (x + y) \overset{\text{ind}}{=} n \cdot x + n \cdot y + (x + y) \overset{(186)}{=} (n + 1) \cdot x + (n + 1) \cdot y. \]

□

**Definition 525.** A **semiring** is a commutative monoid \((R, +)\) with a second associative operation \(\cdot : R \times R \rightarrow R\) called **multiplication**, which extends multiplication with natural numbers. The precise compatibility axioms are listed in definition 525 (a) because they fit nicely into first-order logic (unlike the theory of semimodules, for example, for which we prefer expressing these conditions in the metalogic).

Although not strictly necessary, it will be convenient for us to assume that multiplication has an identity. If a multiplicative identity does not exist, we call \((R, +, \cdot)\) a **nonunital semiring**. A canonical example of a nonunital semiring is a **semiring ideal**. We will not use nonunital semirings, but it is important to acknowledge their existence. In this context, if an identity exists, we will sometimes call \((R, +, \cdot)\) a **unital semiring**.

We call \((R, +)\) the **additive monoid** and \((R, \cdot)\) the **multiplicative monoid** of the semiring. We also consider the **additive group** and the **multiplicative group** as the subsets of invertible elements. Both are instances of proposition 460. The multiplicative group is denoted by \(R^*\); it is discussed further in units.

Semirings have the following metamathematical properties:

(a) The **first-order theory** for semirings extends the **theory of monoids**.

First, we add another infix binary functional symbol \(\cdot\) and a constant 1. The notation for the constant is justified by proposition 530.

We then extend the theory of monoids with **commutativity** for \(+\), **associativity** for \(\cdot\), and the following axioms:

(i) Multiplication on the left distributes over addition:

\[
    \xi \cdot (\eta + \zeta) \equiv \xi \cdot \eta + \xi \cdot \zeta. \tag{186}
\]

(ii) Multiplication on the right also distributes over addition:

\[
    (\xi + \eta) \cdot \zeta \equiv \xi \cdot \zeta + \eta \cdot \zeta. \tag{187}
\]

If multiplication is commutative, right distributivity follows from left distributivity.
(iii) Zero is an absorbing element:

\[ \xi \cdot 0 \equiv 0 \wedge 0 \cdot \xi \equiv 0. \quad (188) \]

(iv) We also restate the identity axiom (165) for the multiplicative unit 1 to highlight its connection with (188):

\[ \xi \cdot 1 \equiv \xi \wedge 1 \cdot \xi \equiv \xi. \quad (189) \]

(b) A first-order homomorphism from the semiring \( R \) to \( T \) is a function \( \varphi : R \to T \) that is a monoid homomorphism both for their additive monoids also for their multiplicative monoids.

(c) The set \( A \subseteq R \) is a submodel of \( R \) if it is a both submonoid of the additive monoid and also of the multiplicative monoid. We call \( A \) a sub-semiring.

As a consequence of proposition 863, the image of a homomorphism \( \varphi : R \to T \) is a sub-semiring of \( A \).

For an arbitrary set \( A \), we denote the generated submodel by \( \langle A \rangle \).

(d) The trivial semiring is the trivial pointed set \( \{0\} \).

See example 527 (a) for some properties of the trivial semiring.

(e) As we shall see in proposition 530, multiplication in \( \cdot \) extends left multiplication with natural numbers in the monoid \((R, +)\). We do have a third operation, however — monoid exponentiation in \((R, \cdot)\).

For any integer \( n \), we have the fundamental property \( 1^n = 1 \).

(f) If multiplication is commutative, we call the semiring itself commutative. Unless multiplication corresponds to function composition, most semirings we will encounter will be commutative.

Notable exceptions to this rule are ordinals. A limit ordinal \( \alpha \), regarded as the set of all smaller ordinals, is a semiring. It is not commutative, however, as shown in example 1074.

(g) Similarly to power set magmas defined in definition 441 (f), the power set \( \text{pow}(R) \) of a semiring is also a semiring with the operations

\[ A \oplus B := \{ x + y \mid x \in A \text{ and } y \in B \} \]
\[ A \odot B := \{ x \cdot y \mid x \in A \text{ and } y \in B \} \]

(h) The corresponding category of \( \mathcal{U} \)-small models \( \mathcal{U}\text{-SRing} \) is concrete over \( \mathcal{U}\text{-CMon} \) with the forgetful functor taking the additive monoids. We denote the category of commutative semirings by \( \text{CSRing} \).
The opposite semiring of \((R, +, \cdot)\) is the semiring \((R, +, \star)\), with multiplication defined as \(x \star y = y \cdot x\).

**Remark 526.** In definition 525, we require semirings to have both an additive identity and a multiplicative identity. This is not consistent with semigroups defined in definition 441 (i), which in general do not have identities.

[GM84, ch. 3] suggest using “diod” (short for “double monoid”) instead of “semiring”. [Gol10, p. xi] describes how the term “diod” may refer to semirings with idempotent addition, i.e. a general form of the tropical semirings defined in definition 528.

We thus prefer using the term “semiring” as we have defined it in definition 525.

**Example 527.** We list several examples of semirings that are not rings.

(a) A semiring is trivial if and only if \(0_R = 1_R\). This follows from (188) and (189).

As a consequence, if \(\varphi : \{0\} \to R\) is a semiring homomorphism, \(R\) is a trivial semiring. This is further strengthened by ?? (UNDEFINED).

(b) The natural numbers are the quintessential example of a semiring. We prove in proposition 8 that they are a semiring.

(c) Every limit ordinal is a monoid under addition, as discussed in example 527 (c), however it is not commutative.

Cardinal addition is commutative, however, and hence for every limit cardinal \(x\), the set of all cardinals smaller than \(x\) is a semiring.

(d) We discussed in example 448 (c) that in a bounded lattice \((X, \lor, \land, \top, \bot)\), both \((X, \lor, \bot)\) and \((X, \land, \top)\) are monoids.

As a consequence of proposition 1257, \(\bot\) is absorbing with respect to \(\land\) and \(\top\) with respect to \(\lor\). Therefore, if the lattice is distributive, as a consequence of proposition 1257, both \((X, \lor, \land)\) and \((X, \land, \lor)\) are semirings.

We refer to these semirings are the positive and negative semiring of the lattice. This terminology comes from example 533 (b).

**Definition 528.** Consider the additive monoid \((\mathbb{N}, +)\) of natural numbers or, more generally, an ordered commutative monoid \((M, +, \leq)\).

We adjoin a top element \(\infty\) to \(M\) that is absorbing with respect to addition. That is, \(x + \infty = \infty\) for every \(x \in M\).

The min-plus semiring over \(M\) is the triple \((M \cup \{\infty\}, \min, +)\). The minimum as a binary operation plays the role of semiring addition, with \(\infty\) as the zero element. The usual addition in \(M\) extended with \(\infty\) plays the role of semiring multiplication, with \(0\) as the multiplicative identity.

We analogously define the max-plus semiring, adjoining a bottom element \(-\infty\) rather than a top element \(\infty\).

We will sometimes use “tropical semiring” to refer to either type of semirings. See remark 529.
Proof of correctness. We will only show distributivity. If \( x \leq y \), since \( \leq \) is compatible with +, we have

\[
\min\{x, y\} + z = x + z \leq y + z. 
\]

Therefore,

\[
\min\{x, y\} + z = \min\{x + z, y + z\}. 
\]

Remark 529. \( \min \)-plus and max-plus semirings are sometimes referred to as the tropical semirings. This term is ambiguous, unfortunately, but it gives rise to the terms “tropical geometry” and “tropical optimization”.

According to [Pin98], the name “tropical semiring” is a dedication to the Brazilian-born Imre Simon. The paper also introduces the terms “tropical integers”, “tropical reals”, etc. [Gol10, p. 3] refers to the more general notion of additively-idempotent semirings. Both reserve the term “tropical semiring” for the case where \( M = \mathbb{N} \). [GM84, ch. 3] does not explicitly use the word “tropical”, but instead refers to semirings as “dioids”, and the latter term sometimes refers to additively-idempotent semirings.

**Proposition 530.** For every semiring, multiplication extends the abelian group multiplication.

More precisely, denote the additive identity by \( 0_R \) and the multiplicative identity by \( 1_R \). Define the following semiring homomorphism:

\[
\iota : \mathbb{N} \rightarrow R \\
\iota(n) := \begin{cases} 
0_R & n = 0, \\
\iota(n - 1) + 1_R & n > 0.
\end{cases}
\]

(190)

This is the unique homomorphism from \( \mathbb{N} \) to \( R \). Furthermore, we have the following analogue to (156):

\[
\iota(n) \cdot x := \begin{cases} 
0_R, & n = 0, \\
\iota(n - 1) \cdot x + x, & n > 1.
\end{cases}
\]

(191)

**Proof.** First note that (191) follows from (190) via right distributivity.

It remains to show that \( \iota \) is a monoid homomorphism, and that it is unique. Clearly \( \iota(0) = 0_R \) and \( \iota(1) = 1_R \). Proving \( \iota(n + m) = \iota(n) + \iota(m) \) and \( \iota(nm) = \iota(n) \cdot \iota(m) \) can be done via nested induction.

Now suppose \( \varphi : \mathbb{N} \rightarrow R \) is a homomorphism. It is clear that \( \varphi(0) = 0_R \) and \( \varphi(1) = 1_R \), and also

\[
\varphi(n + 1) = \varphi(n) + \varphi(1) = \varphi(n) + 1_R.
\]

This implies \( \iota = \varphi \). 

**Proposition 531.** The category of semirings has the following basic properties:

(a) The ring of integers \( \mathbb{Z} \) is an initial object.
(b) The trivial semiring \{0\} is an terminal object.

**Proof.**

**Proof of 531 (a)**. Follows from proposition 530.

**Proof of 531 (b)**. Follows from example 527 (a).

**Definition 532.** An ordered semiring is a commutative semiring \( R \) with a partial order \( \leq \) such that \((R, +)\) is an ordered magma and, additionally, \( x \leq y \) and \( 0 \leq z \) imply \( xz \leq yz \).

As in definition 445, the commutativity condition can be avoided, but then we would need to also require \( zx \leq zy \).

If the semiring is totally ordered, we can use the usual terminology that is conventional for real numbers:

- \( x \) is **positive** if \( x > 0 \).
- \( x \) is **nonnegative** if \( x \geq 0 \).
- \( x \) is **negative** if \( x < 0 \).
- \( x \) is **nonpositive** if \( x \leq 0 \).

**Example 533.** We list several examples of ordered semirings.

(a) The **natural numbers** form an ordered semiring as shown in proposition 11.

(b) We discussed in example 527 (d) that a bounded distributive lattice \((X, \lor, \land)\) can be regarded as a semiring, and so can its opposite lattice.

We discussed in example 446 (c) that both \((X, \lor)\) and \((X, \land)\) are ordered magmas. Both \((X, \lor, \land)\) and \((X, \land, \lor)\) vacuously satisfy the condition from definition 532, which makes them ordered semirings.

All elements of the ordered semiring \((X, \lor, \land)\) are nonnegative and all elements of \((X, \land, \lor)\) are nonpositive. With a slight abuse of notation, we refer to them as the **positive** and **negative** semirings of the lattice.

**Definition 534.** Fix an arbitrary element \( x \) in a semiring. If there exist elements \( l \) and \( r \) such that \( x = lr \), we say that \( l \) is a **left divisor** of \( x \), and that \( r \) is a **right divisor**.

In a commutative semiring, these notions coincide, and we simply use the term “divisor”. If \( x \) is a divisor of \( y \), we write \( x \mid y \) and say that \( y \) is a **multiple** of \( x \). Most rings we will encounter will be commutative, but it is useful to have the weaker notions of left and right divisors.

(a) Divisors of 0 are called **zero divisors**. Due to absorption, every semiring element is a zero divisor. If \( lr = 0 \) for nonzero \( l \) and \( r \), we say that \( l \) (resp. \( r \)) is a **nontrivial** left (resp. right) zero divisor.
(b) Divisors of 1 are called **invertible**, since they are precisely the monoid inverses under multiplication. They are also sometimes called **units**.

The set of all two-sided units of $R$ is precisely the **multiplicative group** $R^\times$.

**Example 535.**

(a) The positive integers are commutative and their left and right divisors coincide. They have no nontrivial zero divisors as a consequence of proposition 8.

(b) A simple example of nontrivial zero divisors is given by the matrix algebra $\mathbb{Z}^{2\times 2}$. We have

\[
\begin{pmatrix}
0 & 1 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix}
= 
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix}.
\]

Therefore, $L$ is a left zero divisor and $R$ is a right zero divisor. The two do not commute because

\[
\begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 \\
0 & 0
\end{pmatrix}
= 
\begin{pmatrix}
0 & 0 \\
1 & 0
\end{pmatrix}.
\]

Nevertheless, $RLRL$ is the zero matrix, so $R$ is a left zero divisor and $L$ is a right zero divisor.

**Proposition 536.** Suppose that $R$ and $S$ are **commutative semirings**.

(a) If $\varphi : R \to S$ is any homomorphism, then $x \mid y$ implies $\varphi(x) \mid \varphi(y)$. The converse holds if $\varphi$ is an isomorphism.

(b) If $R$ and $S$ are isomorphic, the zero divisors of $R$ are precisely the zero divisors of $S$.

(c) If $R$ and $S$ are isomorphic, the units of $R$ are precisely the units of $S$.

**Proof.**

**Proof of 536 (a).** If $x \mid y$, then $xr = y$ for some $r \in R$. Then $\varphi(x)\varphi(r) = \varphi(y)$, hence $\varphi(x) \mid \varphi(y)$. If $\varphi$ is an isomorphism, the converse follows by using $\varphi^{-1} : S \to R$.

**Proof of 536 (b).** Follows from proposition 536 (a) by noting that homomorphisms preserve zeros.

**Proof of 536 (c).** Follows from proposition 536 (a) by noting that homomorphisms preserve ones.

**Proposition 537.** An element of a **commutative semiring** is cancellable if and only if it is not a zero divisor. That is, $x \mid 0$ if and only if $xy = xz$ does not imply $y = z$.

**Proof.** Let $x$ be a nonzero element.
Proof of sufficiency. Suppose that \( x \) is a zero divisor and let \( y \) be such that \( xy = 0 \). For any element \( z \), we have

\[
xy = 0 = x(yz).
\]

But \( y \neq yz \) unless \( z = 1 \). Thus, \( x \) is not cancellable.

Proof of necessity. Suppose that \( x \) is cancellable.

Suppose also that \( xy = 0 \) for some nonzero \( y \). Then \( xy = x0 \), which implies \( y = 0 \). But this contradicts our choice of \( y \).

Thus, \( x \) is not a zero divisor.

Definition 538. We say that the semiring \( R \) is entire if any of the following equivalent conditions hold:

1. \( R \) has no nontrivial zero divisors.
2. \( R \setminus \{0_R\} \) is a cancellative monoid with respect to multiplication.

Proof of correctness. The equivalence follows from proposition 537.

Proposition 539. In an entire commutative semiring, the divisibility relation is a preorder.

It is not a partial order in general. To avoid the nonuniqueness problems described in example 1225, we instead prefer working with ideals. See remark 669 and remark 669 for the general approach.

Proof. Fix a semiring \( R \).

Proof of reflexivity. Clearly every element of \( R \) divides itself.

Proof of transitivity. Let \( x \mid y \mid z \). Then there exist elements \( a \) and \( b \) such that \( y = ax \) and \( z = by \). Hence, \( z = (ba)x \) and \( x \mid z \).

Definition 540. We say that an additive monoid is zerosumfree if the additive group is trivial. That is, if \( x + y = 0 \) implies \( x = y = 0 \).

Example 541. We list several examples of zerosumfree semirings:

(a) By proposition 6, the natural numbers are zerosumfree.

(b) We discussed in example 527 (d) that every bounded distributive lattice \( (X, \lor, \land) \) has two associated semirings.

We will show that the positive semiring \( (X, \lor, \land) \) is zerosumfree. The proof only relies on \( \lor \) being idempotent. Suppose that \( x \lor y = \bot \). Then

\[
\bot = x \lor y \equiv (x \lor x) \lor y \equiv x \lor (x \lor y) = x \lor \bot \equiv x.
\]

Therefore, \( x = \bot \). But \( \bot \lor y = y \), hence \( x \lor y = \bot \) implies \( y = \bot \).

This demonstrates that the positive semiring is zerosumfree.

(c) The min-plus semiring \( (\mathbb{N} \cup \{\infty\}, \min, +) \) discussed in definition 528 is also zerosumfree. Indeed, \( \min \) is idempotent, and the proof is analogous to the one for lattices in example 541 (b).
10.2. Semimodules

Semimodules are generalizations of monoid actions. Notation and terminology-wise, semimodules are somewhat special in that they are very much influenced by linear algebra and analysis, where vector spaces are crucial.

**Definition 542.** Let $X$ be a monoid or, more generally, an object in a category that is concrete over $\text{Mon}$.

Let $\text{end}(X)$ be the endomorphism monoid over $X$. These are necessarily monoid endomorphisms, however they may carry additional structure like being group homomorphisms, semimodule homomorphisms, (semi)lattice homomorphisms or their continuous counterparts.

Define addition in $\text{end}(X)$ pointwise as $[f + g](x) := f(x) + g(x)$. Then $\text{end}(X)$ with pointwise addition and composition is a semiring, which we call the endomorphism semiring over $X$.

**Definition 543.** Fix a semiring $R$, whose elements we will call scalars, and an additive commutative monoid $M$, whose elements we will call vectors. See remark 711 for a discussion of the term “vector”.

We say that $M$ is a semimodule over $R$ if they are compatible in any of the equivalent ways listed below. Analogously to monoid actions, if $R$ is not commutative, we distinguish between left and right semimodules. Rather than “$M$ is a semimodule over $R$”, it is often more convenient to say “$M$ is an $R$-semimodule”.

(a) A left semimodule is a homomorphism from $R$ to the endomorphism semiring $\text{end}(M)$. A right semimodule is a homomorphism from the dual semiring $R^{-1}$ to $\text{end}(M)$.

This definition is concise and natural, but unfortunately not very useful.

(b) The usual way to define a left semimodule is via a binary operation $\cdot : R \times M \to M$ called scalar multiplication that satisfies the following conditions:

(i) Scalar multiplication is a monoid action of the multiplicative monoid $(R, \cdot_R)$ on $M$. The following conditions correspond to (MA1) and (MA2):

\[
1_R \cdot x = x, \quad (r \cdot_R s) \cdot x = r \cdot (s \cdot x). \tag{192}
\]

The second condition can be regarded as a form of associativity.

(ii) Scalar addition distributes over scalar multiplication:

\[
(r +_R s) \cdot x = r \cdot x + s \cdot x. \tag{194}
\]

(iii) Vector addition distributes over scalar multiplication:

\[
r \cdot (x + y) = r \cdot x + r \cdot y. \tag{195}
\]
(iv) The scalar and vector zeros are compatible:

\[ 0_R \cdot x = 0_M = r \cdot 0_M. \] (196)

In practice, we use the same symbol for both scalar and vector addition, and we denote both scalar and vector multiplication via juxtaposition.

Semimodules have the following metamathematical properties:

c) In order to fit the heterogeneous operation \( \cdot \) into the framework of first-order logic models, we can extend the theory of monoids by adding, for every semiring element \( r \), a unary functional symbol \( m_r \). All conditions can then be reformulated via this operation. For example, (193) corresponds to the axiom schema

\[ m_r(\xi) = m_r(m_s(\xi)). \]

d) A first-order homomorphism between two \( R \)-semimodules \( M \) and \( N \) is a function \( \varphi : M \rightarrow N \) that is a monoid homomorphism and satisfies \( \varphi \circ m_r^M = m_r^N \circ \varphi \).

This can be expressed more clearly via the following two conditions, which we call additivity and homogeneity:

\[ \varphi(x + y) = \varphi(x) + \varphi(y) \] (197)

\[ \varphi(rx) = r\varphi(x) \] (198)

Functions satisfying additivity and homogeneity are commonly called linear. These are a central object of study in linear algebra and, to a lesser extent, (linear) functional analysis.

e) The set \( A \subseteq M \) is a submodel of \( M \) if it is a submonoid of \( M \) that is closed under scalar multiplication, i.e. \( rM = m_r[M] \subseteq M \) for every \( r \in R \). We say that \( A \) is an \( R \)-sub-semimodule of \( M \). If \( M \) is a module over some semiring extension \( T \) of \( R \), \( A \) may not be a \( T \)-sub-semimodule. For this reason, we should only use the term “sub-semimodule” of the underlying ring is clear from the context.

As a consequence of proposition 863, the image of an \( R \)-semimodule homomorphism \( \varphi : M \rightarrow N \) is an \( R \)-sub-semimodule of \( M \).

For an arbitrary set \( A \), we denote the generated submodel by span \( A \) and call it the linear span of \( A \).

Remark 555 shows how it is important to be unambiguous about over which semiring we take the span of \( A \). In case of possible ambiguity, we will use subscripts like span\(_R\) \( A \).

The linear span can be characterized via linear combinations — see example 850.

(f) The trivial semimodule is the trivial pointed set \( \{0\} \).
(g) A **bisemimodule** is a triple \((R, A, B)\), where \(A\) is a left \(R\)-semimodule, \(B\) is a right \(R\)-semimodule, and the following associativity condition holds for \(a \in A\), \(r \in R\) and \(b \in B\):

\[
(a \cdot_A r) \cdot_B b = a \cdot_A (r \cdot_B b).
\]  

(199)

(h) For a fixed semiring \(R\), the category of \(\mathcal{U}\)-small models \(\mathcal{U} \text{-SMO}_{R}\) of left semimodules is **concrete** over \(\mathcal{U} \text{-MON}\).

Other notations are in use, for example \(R - \text{Mod}\) in [Alu09, p. 158], that better highlight whether we are considering left or right semimodules. We will prefer \(\text{Mod}^{\text{op}}_R\) for the category of right modules.

**Proof.**

**Proof that 543 (a) implies 543 (b).** Fix a semiring homomorphism \(\varphi : R \to \text{end}(M)\) and define the operation \(r \cdot x := \varphi(r)(x)\).

We will verify that all conditions from **definition 543 (b)** hold for this operation.

- By definition, \(\varphi\) is a monoid action of \((R, \cdot)\) on \((M, \circ)\).
- Distributivity of scalar addition holds because \(\varphi\) is a **magma homomorphism** from \((R, +)\) to \((M, +)\).
- Distributivity of vector addition holds because, for each \(r\), \(\varphi(r)\) is a magma endomorphism of \((M, +)\).
- Since \(\varphi\) is a monoid homomorphism from \((R, +)\) to \((R, \cdot)\), it preserves identities and hence

\[
0_R \cdot x = \varphi(0_R)(x) = [y \mapsto 0_M](x) = 0_M.
\]

This proves half of **definition 543 (b iv)**.

- Since, for each \(r\), \(\varphi(r)\) is a monoid endomorphism of \((M, +)\), we have

\[
r \cdot 0_M = \varphi(r)(0_M) = 0_M.
\]

This proves the other half of **definition 543 (b iv)**.

**Proof that 543 (b) implies 543 (a).** Let \(\cdot : R \times M \to M\) be an operation satisfying all conditions from **definition 543 (b)**. Define the function \(\varphi(r) := (x \mapsto r \cdot x)\). We will show that this is a semiring homomorphism.

It preserves both identities because

\[
\varphi(0_R) = (x \mapsto 0) = 0_{\text{end}(M)}
\]

and

\[
\varphi(1_R) = (x \mapsto x) = \text{id}_M.
\]

263
We must also show that it preserves both binary operations. Clearly
\[ \varphi(r + s) = (x \mapsto (r + s)x) \equiv (x \mapsto rx + sx) = (x \mapsto rx) + (x \mapsto sx) = \varphi(r) + \varphi(s). \]

For multiplication, we have
\[ \varphi(rs) = (x \mapsto (rs)x) \equiv (x \mapsto r(sx)) = \left( x \mapsto \varphi(r)\left(\varphi(s)(x)\right) \right) = \varphi(r) \circ \varphi(s). \]

\[ \square \]

**Proposition 544.** Semimodules have the following basic properties:

(a) The union of a monotone sequence

\[ N_1 \subseteq N_2 \subseteq \cdots \]

of \( R \)-sub-semimodules of \( M \) is also an \( R \)-sub-semimodule of \( M \).

**Proof.**

**Proof of 544 (a).** Trivial.

\[ \square \]

**Proposition 545.** Every semiring is a bisemimodule over itself with scalar multiplication given by the semiring multiplication.

This result specializes to proposition 566.

**Proof.** Fix a semiring \( R \). We will show that \( \cdot \) satisfied the conditions in definition 543 (b).

- The identity law (192) holds because 1 is a multiplicative identity of \( M \).
- The associativity-like law (193) follows from associativity of multiplication.
- The two distributivity laws (194) and (195) follow from left and right distributivity on \( R \).
- The absorption law (196) follows from absorption on semirings.

All the above also hold for right semimodules rather than left.

\[ \square \]

**Proposition 546.** The categories \( \text{CMon} \) of commutative monoids and \( \text{SMod}_\mathbb{N} \) of natural number semimodules are isomorphic.

More concretely, every commutative monoid \( M \) is a left semimodule over \( \mathbb{N} \) with scalar multiplication given by recursively defined multiplication

\[ n \cdot x := \begin{cases} 0_M, & n = 0, \\ n \cdot x + x, & n > 1. \end{cases} \quad (200) \]

Conversely, in every semimodule over \( \mathbb{N} \), scalar multiplication matches the recursively defined multiplication.

This result specializes to proposition 567 and proposition 605.
Proof.

Proof of sufficiency. Let $M$ be a commutative monoid. The operation $\cdot : \mathbb{N} \times M \to M$ defined in proposition 530 satisfies the conditions in definition 543 (b) as either a direct consequence of the definition or as a consequence of proposition 524. The homomorphisms are thus also compatible.

Proof of necessity. Let $M$ be a semimodule over $\mathbb{N}$. We will use induction to show that the? holds.

- For $n = 0$, this follows from the absorption law (196).
- If $n \cdot x = n \cdot x + x$, then by scalar distributivity, $(n + 1) \cdot x = n \cdot x + 1 \cdot x$. The multiplicative identity law (192) then shows that $1 \cdot x = x$, which concludes our proof.

The homomorphisms are thus also compatible.

Proposition 547. For a set $A$ and an $R$-semimodule $N$, the set $\text{fun}(A, N)$ of all functions from $A$ to $N$ is itself an $R$-semimodule with the following operations:

(a) Pointwise addition $[f + g](x) := f(x) + g(x)$

(b) Pointwise scalar multiplication $[t \cdot f](x) := t \cdot f(x)$

If $A$ is also an $R$-algebra, we denote the semimodule of all $R$-linear maps by $\text{hom}(A, N)$. This extends to proposition 568.

Proof. By proposition 866, $N$ is an $R$-semimodule.

Definition 548. The support of a function $f : S \to R$ from any set $S$ to a semiring $R$ is the set $\text{supp}(f) := \{x \in S \mid f(x) \neq 0_R\}$.

Definition 549. The direct product of a family of $R$-semimodules $\{M_k\}_{k \in \mathcal{K}}$ is their monoid direct product $\prod_{k \in \mathcal{K}} M_k$ with the additional componentwise scalar product $r \cdot \{x_k\}_{k \in \mathcal{K}} := \{r \cdot x_k\}_{k \in \mathcal{K}}$.

As in the case of general monoids, the direct sum $\bigoplus_{k \in \mathcal{K}} M_k$ is the submonoid of the direct product consisting only of tuples with finite support. That is, tuples with only finitely many nonzero components.

Proposition 550. We present a refinement to proposition 451.

(a) The categorical product of the family $\{M_k\}_{k \in \mathcal{K}}$ in the category $\text{SMod}$ of semimodules is their direct product $\prod_{k \in \mathcal{K}} M_k$. 

265
(b) The categorical coproduct of the family \(\{M_k\}_{k\in\mathcal{K}}\) in the category \(\text{CSMon}\) of commutative semimodules is their direct sum \(\bigoplus_{k\in\mathcal{K}} M_k\).

**Proof.** Simple refinement of proposition 451. \(\Box\)

**Definition 551.** Fix a semiring \(R\). We associate with every set \(A\) its free \(R\)-semimodule \(R^\oplus A\) over \(R\) defined as the set

\[
R^\oplus A := \bigoplus_{x \in A} R = \{t : A \to R \mid t \text{ has finite support}\}.
\]

In the case when \(R\) is the semiring \(\mathbb{N}\) of natural numbers, \(\mathbb{N}^\oplus A\) is the set of finite multisets over \(S\).

We call \(A\) the **basis** of \(R\). As with general direct sums, we regard the function \(t\) as the indexed family \(\{t_x\}_{x \in A}\), and we call the indexed family a **linear combination** over \(R\).

By proposition 866, \(R^\oplus A\) inherits addition and multiplication from \(R\) and is actually a semiring. Scalar multiplication can be defined as

\[
\cdot : R \times R^\oplus A \to R^\oplus A, \\
\cdot : (r, \{t_x\}_{x \in A}) \mapsto \{r \cdot t_x\}_{x \in A}.
\]

Finally, define the canonical inclusion.

\[
t_A : A \to R^\oplus A, \\
t_A(x) := \left\{\begin{array}{ll}
1_R, & y = x \\
0_R, & y \neq x
\end{array}\right.
\]

Free right semimodules require trivial adjustments.

**Theorem 552** (Free semimodule universal property). Fix a semiring \(R\) and a set \(A\). The free \(R\)-semimodule \(R^\oplus A\) over \(R\) is the unique up to a unique isomorphism semimodule that satisfies the following universal mapping property:

For every semimodule \(M\) over \(R\) and every function \(e : A \to M\), there exists a unique \(R\)-semimodule homomorphism \(\Phi_e : R^\oplus A \to M\) such that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{e} & M \\
\downarrow t_A & & \downarrow \Phi_e \\
R^\oplus A
\end{array}
\]

(201)

Via remark 1198, \(A \mapsto R^\oplus A\) becomes left adjoint to the forgetful functor

\[
U : \text{SMod}_R \to \text{Set}.
\]

The function \(e\) assigns a value from \(M\) for each component of a linear combination, while \(\Phi_e\) evaluates the obtained sum. We call \(\Phi_e\) the **linear extension** or **linear combination evaluation map** corresponding to \(e\).
Proof. For every function $e : A \to M$, we want
\[ \Phi_e(x) = e(x). \]
This suggests the definition
\[ \Phi_e : R^{\oplus A} \to M, \]
\[ \Phi_e(t_x x \in A) := \sum_{x \in A} t_x \cdot e(x). \]

We discuss well-definedness of infinitary operations in direct sums in remark 794 (c iii). □

Remark 553. The linear combination $\sum_{x \in A} t_x x$ can instead be written as $\sum_{k=1}^{n} t_k x_k$, where $x_1, ..., x_n$ are the values in $A$ for which the scalars $t_1, ..., t_n$ are nonzero (we denote $t_{x_k}$ by $t_k$ for brevity). This is actually the dominating convention, although we sometimes use the former notation and. In vector spaces, we utilize the projection functionals from definition 621.

This issue is discussed in more generality in remark 794 (c iii).

Proposition 554. For a set $A$ in an $R$-semimodule $M$, the linear span of $A$, defined as the $R$-submodule generated by $A$ in the sense of definition 852, equals the set of all linear combinations over $A$.

We say that $A$ is a generating set of $M$.

Compare this result to proposition 575 for algebras.

Proof. Figure 21 shows a syntax tree for a given linear combination, which can be traversed and evaluated to obtain a vector in $M$. Furthermore, it must be a member of span $S$ since the latter is closed under vector addition and scalar multiplication with members of $S$. Hence, the set $L$ of all linear combinations over $S$ is a subset of span $S$.

Generalizing the syntax tree construction from fig. 21, we see that $L$ satisfies definition 848 (b), and is thus a submodule of $M$. Since span $S$ is the smallest module containing $S$, we have span $S = L$.

\[
\begin{align*}
(ax + by) + cz \\
ax + by \\
ax \\
by
\end{align*}
\]

Figure 21: A linear combination is simply a superposition of scalar multiplication and binary addition.

□
Remark 555. If $M$ is both an $R$-semimodule and a $T$-semimodule, proposition 554 highlights a fundamental difference between the generated $R$-sub-semimodule and the generated $T$-sub-semimodule.

For example, the $\mathbb{N}$-sub-semimodule generated by 2 is the semiring $2\mathbb{N}$ of even natural numbers, while the $\mathbb{R}_{\geq 0}$-sub-semimodule generated by 2 is $\mathbb{R}_{\geq 0}$ itself.
10.3. Semiring ideals

When regarding semirings as semimodules over themselves, as per proposition 545, we obtain the important notion of ideals.

**Definition 556.** Fix a semiring $R$. A **left ideal** of $R$ is a sub-semimodule of $R$ when regarded as a left semimodule over itself, and a **right ideal** is defined analogously.

If $I$ is both a left and right ideal of $R$, we say that it is a **two-sided ideal** or simply **ideal**. More explicitly, $I$ is a two-sided ideal of $R$ if it is a submonoid of the additive monoid of $R$ that is closed under left and right multiplication, i.e. $RI = IR$.

When **quotient rings** are involved, we will have no choice but to work with two-sided ideals. If multiplication is commutative, every left ideal is a right ideal and there is no distinction between the two. Otherwise, we will usually consider left ideals by default. Right ideals are left ideals in the opposite semiring, and thus we lose no generality.

(a) For an arbitrary subset $A$ of $R$, we call the (left) **linear span** of $A$ the left ideal generated by $A$. Explicitly, this is the set

$$\sum_{a \in A} Aa = \left\{ \sum_{k=1}^{n} t_k a_k \mid n > 0 \text{ and, for } k < n, t_k \in R \text{ and } a_k \in A \right\}.$$ 

If $A = \{a_1, \ldots, a_n\}$, we say that the ideal is **finitely generated** and write

$$Aa_1 + \cdots + Aa_n.$$ 

For right ideals, this becomes

$$a_1 A + \cdots + a_n A.$$ 

This is discussed in proposition 575 for the case of commutative rings, where we use the notation $\langle A \rangle$. In general rings, we are more explicit for the sake of avoiding possible confusion.

(b) An ideal generated by a single element is called a **principal ideal**. In a general ring, there can be left, right and two-sided principal ideals.

(c) The **product ideal** $IJ$ of $I$ and $J$ is

$$\left\{ \sum_{k=1}^{n} i_k j_k \mid n > 0 \text{ and, for } k < n, i_k \in I \text{ and } j_k \in J \right\}.$$ 

This notation is unfortunately inconsistent with the pointwise product $\{ij \mid i \in I, j \in J\}$ from definition 441 (f); it is actually the ideal generated by the pointwise product.

(d) If $P$ is a proper ideal and if from $IJ \subseteq P$ it follows that $I \subseteq P$ or $J \subseteq P$ (or both), we say that $P$ is a **prime ideal**.

When working with commutative semirings, proposition 558 (h) is instead sometimes taken as the definition of a prime ideal.
(e) If $I + J = R$ for proper ideals $I$ and $J$, we say that $I$ and $J$ are coprime. Equivalently, $I$ and $J$ are coprime if their sum contains a unit.

(f) A (left) maximal ideal is a proper (left) ideal that is maximal with respect to set inclusion. The maximal ideals are the predecessors of $R$ in the lattice of (left) ideals described in Proposition 559.

Remark 557. A proper semiring ideal is a canonical example of a nonunital sub-semiring. As a consequence of Proposition 558 (a), a proper ideal cannot contain the multiplicative identity 1, and is thus not a sub-semiring unless we allow sub-semirings to not contain 1.

Proposition 558. The left ideals of a semiring $R$ have the following basic properties:

(a) An ideal contains a unit if and only if it is not proper. In particular, $R = \langle 1_R \rangle$.

(b) A semiring element is a unit if and only if it does not belong to any proper ideal.

(c) We have $\langle x \rangle \subseteq \langle y \rangle$ if and only if $y \mid x$ for two-sided ideals and two-sided divisors.

More generally, we have $Rx \subseteq Ry$ if and only if $y$ is a right divisor of $x$. Note how $Rx$ and $Ry$ are left principal ideals but $y$ is a right divisor.

(d) The union of a monotone sequence

$$I_1 \subseteq I_2 \subseteq \cdots$$

if ideals is again an ideal.

(e) Every (left or right) maximal ideal is prime.

Proposition 683 (c) is a converse that holds for principal ideal domains.

(f) We have $IJ \subseteq I \cap J$. The converse inclusion holds if $R$ is commutative and if $I$ and $J$ are coprime.

The following require $R$ to be commutative:

(g) In a commutative semiring, the product of the principal ideals $\langle x \rangle$ and $\langle y \rangle$ is $\langle xy \rangle$. 
(h) In a commutative semiring, an equivalent condition to \( P \) being prime is that \( xy \in P \) implies \( x \in P \) or \( y \in P \) (or both).

(i) In a commutative semiring, every prime ideal is radical.

Proof.

**Proof of 558 (a).** We will prove that there exists a unit \( u \in I \) if and only if \( I = R \).

**Proof of sufficiency.** Let \( u \in I \) be a unit. Then \( 1_R = u^{-1}u \in I \). It follows that \( 1_R \cdot x = x \) for any \( x \in R \), thus \( IR = I \). But \( I \) is an ideal, hence we have that \( IR = I \). Therefore, \( I = IR = R \).

**Proof of necessity.** If \( I = R \), then obviously \( 1_R \in I \).

An analogous proof follows for the case when \( I \) is a right ideal.

**Proof of 558 (b).**

**Proof of sufficiency.** Suppose that \( x \) is a unit and that \( x \) belongs to some proper ideal \( I \). Then \( Rx = R \), implying that \( I = R \), which is a contradiction.

**Proof of necessity.** Suppose that \( x \) does not belong to any proper ideal. Then \( Rx \) is not a proper ideal, implying that \( R = Rx \). There exists some \( y \) such that \( yx = 1_R \). Hence, \( x \) is a unit.

**Proof of 558 (c).**

**Proof of sufficiency.** Suppose that \( Rx \subsetneq Ry \). Then \( x \in Ry \), and hence there exists an element \( l \) of \( R \) such that \( x = ly \). So \( y \) is a right divisor of \( x \).

**Proof of necessity.** Suppose that \( y \) is a right divisor of \( x \). Then there exists an element \( l \) of \( R \) such that \( x = ly \). Thus, \( x \in Ry \), and hence \( Rx \subsetneq Ry \).

**Proof of 558 (d).** Follows from proposition 544 (a).

**Proof of 558 (e).** Let \( M \) be a maximal left ideal and let \( IJ \subseteq M \). Suppose that both \( M \setminus I \) and \( M \setminus J \) are nonempty.

Then there exist elements \( i \in I \), \( j \in J \), \( m_i \in M \) and \( m_j \in M \) such that

\[
i + m_i = j + m_j = 1_R.
\]

Then

\[
1_R = (i + m_i)(j + m_j) = 
\begin{bmatrix}
\begin{array}{c}
ij \\
M
\end{array}\end{bmatrix} + 
\begin{bmatrix}
\begin{array}{c}
m_i j \\
M
\end{array}\end{bmatrix} + 
\begin{bmatrix}
\begin{array}{c}
m_i m_j \\
M
\end{array}\end{bmatrix}.
\]

Hence,

\[
1_R = (i + m_i)(j + m_j) \in M,
\]

which contradicts our assumption that \( M \) is maximal.

Therefore, \( M \setminus I \) and \( M \setminus J \) cannot both be nonempty.
Proof of 558 (f). We will first show that $IJ \subseteq I \cap J$. Let
\[ \sum_{k=1}^{n} x_k y_k \in IJ. \]
For each $k$, $x_k y_k$ belongs to both $I$ and to $J$. Hence, the sum over $k$ also belongs to the intersection. Therefore,
\[ IJ \subseteq I \cap J. \]

Proof of 558 (g).
Proof of sufficiency. Let $z \in \langle x \rangle \langle y \rangle$. Then there exist elements $x_x$ of $\langle x \rangle$ and $y_y$ of $\langle y \rangle$ such that $z = x_x y_y$, and elements $r_x$ and $r_y$ of $R$ such that $xr_x = x_x$ and $yr_y = y_y$.
Therefore,
\[ z = (xr_x)(yr_y) \in \langle xy \rangle. \]

Proof of necessity. Let $z \in \langle xy \rangle$. Then there exists an element $r$ of $R$ such that $z = rxy = (rx)(y)$, hence $z \in \langle x \rangle \langle y \rangle$.

Proof of 558 (h). Suppose that $R$ is commutative.
Proof of sufficiency. Let $P$ be prime and let $xy \in P$. Then $\langle xy \rangle \subseteq P$. By proposition 558 (g), $\langle x \rangle \langle y \rangle \subseteq P$, and hence $\langle x \rangle \subseteq P$ or $\langle y \rangle \subseteq P$. Therefore, $x \in P$ or $y \in P$.

Proof of necessity. Let $P$ be an ideal such that $xy \in P$ implies $x \in P$ or $y \in P$. Let $IJ \subseteq P$ and suppose that there exist $i \in I \setminus P$ and $j \in J \setminus P$.
Obviously $ij \in I$. But since $P$ is prime, it follows that $i \in P$ or $j \in P$.
The obtained contradiction shows that $I$ or $J$ must be a subset of $P$. Therefore, $P$ is prime.

Proof of 558 (i). Suppose that $R$ is commutative, let $P$ be a prime ideal and let $x^n \in P$ for $n > 0$. We will show that $x \in P$.
We proceed via induction on $n$. The case $n = 1$ is trivial. Suppose that $x^{n-1} \in P$ implies $x \in P$, and let $x = x \cdot x^{n-1} \in P$. Since $P$ is prime and $R$ is commutative, by proposition 558 (h), $x \in P$ or $x^{n-1} \in P$. In the latter case, we use the inductive hypothesis to show that $x \in P$.
Generalizing on $x$, we conclude that $P$ is a radical ideal. \qed

Proposition 559.
(a) The set $\mathcal{I}$ of all ideals of a semiring $R$ is itself a semiring with the set addition and multiplication from power set semiring $\text{pow}(R)$.

(b) Furthermore, $\mathcal{I}$ is an ordered semiring with respect to set inclusion.

(c) The supremum of $I$ and $J$ is their sum $I + J$ and their infimum is their intersection $I \cap J$.
With this, $\mathcal{I}$ becomes a lattice.

Proof.
Proof of 559 (a). Associativity and commutativity in $\mathcal{I}$ are inherited from $\mathcal{R}$, as well as both left and right distributivity. Distributivity ensures that $I + J$ is an ideal, while associativity of multiplication ensures that $IJ$ is an ideal.

Proof of 559 (b). We must now prove that the partial order $\subseteq$ is compatible with addition and multiplication. Suppose that $I \subseteq J$ and let $H$ be any ideal in $\mathcal{J}$. Then

$$I + H \subseteq J + H$$

and

$$IH \subseteq JH.$$

Therefore, $J$ is an ordered semiring.

Proof of 559 (c). Since $0 \in I$, obviously $I \subseteq I + J$, and thus $I + J$ is an upper bound of $I$ and $J$. If $H$ is any other upper bound, it must contain the sums of all elements of $I$ and all elements of $J$, hence $I + J \subseteq H$. Therefore, $\sup\{I, J\} = I + J$.

For $I \cap J$, it is an infimum of $I$ and $J$ as a consequence of proposition 938 (c).

Example 560. We list several examples of semiring ideals.

(a) The simplest example of an ideal that is not principal is the ideal $\langle 2, 3 \rangle$ in $\mathbb{N}$.

To see that it is not principal, suppose that $\langle n \rangle = \langle 2, 3 \rangle$ for some natural number $n$. This implies that there exist nonzero numbers $a$ and $b$ such that $n = 2a + 3b$. Hence, $n > 2a > a$ and $n > 3b > b$. But then neither 2 nor 3 belongs to $\langle n \rangle$, contradicting our assumption.

(b) The zero ideal $\langle 0 \rangle$ in $\mathbb{N}$ is prime but not maximal.

Indeed, since $\mathbb{N}$ is entire, $\langle 0 \rangle = \{0\}$ and thus $\langle 0 \rangle$ is prime. But it is not maximal since it is contained in every other ideal.

(c) For natural numbers, $\langle n \rangle = \langle m \rangle$ implies $n = m$.

Indeed, by proposition 558 (c), $n \mid m$ and $m \mid n$. Thus, there exist numbers $a$ and $b$ such that $n = am$ and $m = bn$, hence $n = abn$. Since the semiring $\mathbb{N}$ is entire, we can cancel $n$ to obtain $ab = 1$. Then $a = b = 1$, and hence $n = m$.

(d) A natural number $n$ is prime if and only if $\langle n \rangle$ is a prime ideal in $\mathbb{N}$.

Suppose that $n$ is prime and let $n \mid mk$. From lemma 26 (Euclid’s lemma) it follows that either $n \mid k$ or $n \mid m$, hence proposition 558 (h) is satisfied and $\langle n \rangle$ is a prime ideal.

In the other direction, suppose that $\langle n \rangle$ is a prime ideal and let $n = ab$. By proposition 558 (g), $\langle n \rangle = \langle a \rangle \langle b \rangle$. Since $\langle n \rangle$ is a prime ideal, $\langle a \rangle \subseteq \langle n \rangle$ or $\langle b \rangle \subseteq \langle n \rangle$.

Therefore, $\langle n \rangle = \langle a \rangle$ or $\langle n \rangle = \langle b \rangle$. By example 560 (c), $n = a$ or $n = b$, which in turn implies that the other is a unit.

Therefore, $n$ is a prime number.

273
(e) Consider the matrix algebra $\mathbb{Z}^{2 \times 2}$. The set
\[
\left\{ \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} \mid a, b \in \mathbb{Z} \right\}.
\]
is a left ideal. It is not a right ideal, however, because
\[
\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.
\]

(f) Consider the bivariate polynomial semiring $\mathbb{N}[X, Y]$ over natural numbers. Since
\[(X + Y)^2 = X^2 + 2XY + Y^2,\]
we have
\[
\langle X^2 + 2XY + Y^2 \rangle \subseteq \langle X + Y \rangle.
\]

(g) Ideals in polynomial semirings are often studied, but we can also study polynomials in ideal semirings, i.e. polynomials over the semiring $\mathcal{I}$ of ideals of a semiring $R$. For example,
\[
I^2 J + K
\]
is a trivariate polynomial function over $\mathcal{I}$.

(h) If $M$ is a maximal ideal and $x \in R \setminus M$, then $M$ and $\langle x \rangle$ are coprime.

[Theorem 561](Gol10 prop. 6.59) (Maximal ideal theorem). Every proper semiring ideal is contained in a maximal ideal.

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (l).

Proof. We will discuss equivalence with theorem 1240 (Zorn’s lemma).

**Proof that Zorn’s lemma implies maximal ideal theorem.** Let $I$ be a proper ideal in the semiring $R$. Denote by $\mathcal{I}$ the set of all proper ideals in $R$ that contain $I$. The union of every chain in $\mathcal{I}$ is again an ideal, and by Zorn’s lemma, $\mathcal{I}$ has a maximal element. That is, there exists a maximal ideal in $\mathcal{I}$ that contains $I$.

**Proof that maximal ideal theorem implies Zorn’s lemma.** In [Hod79], Hodges proves that the statement “every unique factorization domain has a maximal ideal” implies Zorn’s lemma. We have an even stronger antecedent.

**Definition 562.** Let $I$ be an ideal in the commutative ring $R$. The **radical** of $I$ is a specific ideal containing $I$ that we will define shortly. A **radical ideal** is an ideal that is equal to its radical. The **nilradical** of the ring $R$ is radical $\sqrt{(0_R)}$ of the zero ideal, whose elements we call **nilpotent elements**.

The radical of the ideal $I$ is the ideal defined equivalently through any of the following:

\[ \sqrt{I} := \{ x \in R \mid \exists n \in \mathbb{Z}_{>0} \cdot x^n \in I \} \]  

(202)
(b) $\sqrt{I}$ is the intersection of all prime ideals of $R$ containing $I$.

**Proof.**

**Proof of correctness of (202).** We only need to prove that the radical $\sqrt{I}$ of the ideal $I$ is an ideal.

Multiplicative closure is simpler. If $x$ belongs to $\sqrt{I}$, then there exists a power $x^n$ that belongs to $I$. Let $r$ be any member of $r$. Then $rx = rx^n \in I$ since $I$ is closed with respect to multiplication.

Additive closure is a bit more involved. If $x$ and $y$ both belong to $\sqrt{I}$, then there exist powers $n$ and $m$ such that $x^n \in I$ and $y^m \in I$. Let $u := n + m - 1$. By theorem 1287 (Newton’s binomial theorem),

$$(x + y)^u = \sum_{k=0}^{u} \binom{u}{k} x^k y^{u-k}.$$  

- If $k < n$, then $x^k y^{u-k} = (x^k y^{n-k-1})y^m$ and, since $y^m \in I$, we have $x^k y^{u-k} \in I$.
- If $k \geq n$, then $x^k y^{u-k} = x^n(x^{k-n} y^{u-k})$ and, since $x^n \in I$, we have $x^k y^{u-k} \in I$.

Since $I$ is closed under addition, $(x + y)^u \in I$.

**Proof that 562 (a) implies 562 (b).** Let $x \in \sqrt{I}$. That is, there exists a positive integer $n$ such that $x^n \in I$. Let $P$ be a prime ideal containing $I$. We will show that $x \in P$.

Since $P$ is prime, by proposition 558 (h), $x^{n-1} \in P$ or $x \in P$. If $x^{n-1} \in P$, then $x^{n-2} \in P$ or $x \in P$. Proceeding by induction on $k$ in $x^{n-k}$, we eventually obtain that $x \in P$.

**Proof that 562 (b) implies 562 (a).** Conversely, let $x$ be a member of every prime ideal containing $P$. We will show that $x \in \sqrt{I}$.

Suppose that $x^n \not\in I$ for every $n$ and consider the following family of ideals:

$$\mathcal{H} := \{ J \text{ is an ideal of } R \text{ containing } I \mid \forall n \in \mathbb{Z}_{>0} \cdot x^n \not\in J \}.$$  

It is nonempty because $I \in \mathcal{H}$.

For every chain of ideals in $\mathcal{H}$, their union is also an ideal in $\mathcal{H}$. By theorem 1240 (Zorn’s lemma), $\mathcal{H}$ has a maximal element $H$. We will show that $H$ is prime.

From $AB \subseteq H$ it follows that $AB \in \mathcal{H}$. If we suppose that neither $A$ nor $B$ belongs to $\mathcal{H}$, we obtain that there exist positive integers $n$ and $m$ such that $x^n \in A$ and $x^m \in B$. But $x^{n+m}$ must then belong to $AB$, which contradicts $AB \in \mathcal{H}$. The obtained contradiction demonstrates that $H$ is a prime ideal. But this is impossible since, by assumption $x$ is contained in every prime ideal, and $x \not\in H$. So this must contradict our previous assumption that $x^n \neq 0$ for every $n$.

Therefore, $x \in \sqrt{I}$.  

**Example 563.** We list examples of radical ideal.
Suppose that the natural number \( m \) has a prime decomposition

\[
m = p_1^{k_1} \cdots p_n^{k_n}.
\]

Then the radical of its principal ideal is

\[
\sqrt{\langle p_1^{k_1} \cdots p_n^{k_n} \rangle} = \langle p_1 \cdots p_n \rangle.
\]

Indeed, for any \( a p_1 \cdots p_n \) from the radical, with \( k := \max\{k_1, \ldots, k_n\} \) we have

\[
(a p_1 \cdots p_n)^k = (a p_1^{k-k_1} \cdots p_n^{k-k_n}) p_1^{k_1} \cdots p_n^{k_n}.
\]

To obtain \( m \), we can take \( k = 1 \) and \( a = p_1^{k_1-1} \cdots p_n^{k_n-1} \):

\[
ap_1 \cdots p_n = p_1^{k_1} \cdots p_n^{k_n} = m.
\]

Particular examples of this are

- The ideal \( \langle 6 \rangle \) is radical. It is not prime since \( \langle 2 \rangle \langle 3 \rangle = \langle 6 \rangle \), but neither are subsets.
- For any prime \( p \), \( \langle p \rangle \) is radical
- For any prime power \( p^n \), \( \sqrt{\langle p^n \rangle} = \langle p \rangle \). For example, \( \sqrt{\langle 4 \rangle} = \langle 2 \rangle \).

Consider the matrix ring \( \mathbb{N}^{2 \times 2} \). The matrix

\[
A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}
\]

is a nilpotent element of \( \mathbb{N}^{n \times n} \) because \( A^2 \) is the zero matrix. We refer to such matrices as \textbf{nilpotent matrices}.

The transposed matrix \( A^T \) is also nilpotent. Hence, their linear combinations are also nilpotent.
10.4. Algebras over semirings

Algebras are usually defined for fields or at least commutative rings. We extend this to semirings for the purposes of polynomial semirings.

**Definition 564.** Generalizing *linear maps*, if $M_1, \ldots, M_n$ and $N$ are $R$-modules, we say that the function

$$f : M_1 \times \cdots \times M_n \rightarrow N$$

is *multilinear* (*bilinear* for $n = 2$) if it is linear in each component. That is, for every tuple

$$(x_1, \ldots, x_n) \in M_1 \times \cdots \times M_n,$$

and for every index $k = 1, \ldots, n$, the following function is linear:

$$y \mapsto f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n)$$

**Definition 565.** An *algebra* over a commutative semiring $R$ is an $R$-semimodule $M$ with an associative bilinear vector multiplication operation. This makes $M$ a nonunital ring. By default, we will also assume that $M$ has a multiplicative unit, although nonunital algebras are just as valid as nonunital rings.

As in the case of general rings, by “$M$ is commutative”, we will mean that vector multiplication is commutative. Furthermore, although we assume it by default, if needed, we will distinguish between associative and non-associative algebra.

We identify every element $t$ of $R$ with its canonical embedding $t \cdot 1_M$ in $M$, and thus we can also regard $R$ as a sub-semiring of $M$.

Algebras have the following metamathematical properties:

(a) The *first-order theory* for algebras extends the *theory of commutative semimodules*. We add a new *infix* binary function symbol $\odot$ to the language, and add to the theory all semiring axioms from *definition 525* (a) for $+$ and $\odot$. We must also add axioms ensuring that $\odot$ is bilinear. Additivity follows from distributivity, hence it remains to account for homogeneity. Using the notation of *definition 543* (c), this amounts to the following axiom schemas:

$$m_r(x) \odot y = m_r(x \odot y),$$

$$x \odot m_r(y) = m_r(x \odot y).$$

(b) A *first-order homomorphism* between two $R$-algebras $M$ and $N$ is a linear map that also preserves vector multiplication.

(c) The set $A \subseteq M$ is a *submodel* of $M$ if it is a *submodule* of $M$ that is closed under algebra multiplication. We say that $A$ is a *subalgebra*.

As for general submodules, *remark 555* shows how it is important to be unambiguous about over which semiring we consider the subalgebra.

As a consequence of *proposition 863*, the image of an $R$-algebra homomorphism is a subalgebra of its range.
(d) The trivial semimodule is the trivial pointed set \{0\}.

(e) We denote the category of algebras over \(R\) by \(\text{Alg}_R\) and the subcategory of commutative algebras by \(\text{CAlg}_R\).

**Proposition 566.** Every semiring \(R\) is an \(R\)-algebra with both scalar and vector multiplication given by the multiplication in \(R\).
This extends proposition 545.

**Proof.** Follows from proposition 545 by noting that bilinearity follows from left distributivity in \(R\). \(\square\)

**Proposition 567.** The categories \(\text{SRing}\) of semirings and \(\text{Alg}_\mathbb{N}\) of natural number algebras are isomorphic.
Compare this result to proposition 546 and proposition 648.

**Proof.** Follows from proposition 546 by noting that, as in the proof of proposition 566, distributivity implies bilinearity. \(\square\)

**Proposition 568.** For a set \(A\) and an \(R\)-algebra \(N\), the set of all functions from \(A\) to \(N\) is itself an \(R\)-algebra with the following operations:

(a) **Pointwise addition**
\[
[f + g](x) := f(x) + g(x)
\]

(b) **Pointwise scalar multiplication**
\[
[t \cdot f](x) := t \cdot f(x)
\]

(c) **Pointwise vector multiplication**
\[
[f \odot g](x) := f(x) \cdot g(x)
\]

In practice, we use juxtaposition \(fg\) or \(f \cdot g\) instead of \(f \odot g\).

If \(A\) is also an \(R\)-algebra, we denote the set of all \(R\)-algebra homomorphisms by \(\text{hom}(A, N)\). This result extends proposition 547.

**Proof.** By proposition 866, \(N\) is both an \(R\)-semiring and an \(R\)-semimodule. Compatibility comes from left distributivity in \(N\). \(\square\)

**Definition 569.** A **multi-index** over the plain set \(\mathcal{K}\) is a member of the free \(\mathbb{N}\)-semimodule \(\mathcal{K}^{\otimes \mathbb{N}}\) over \(\mathcal{K}\). We endow \(\mathcal{K}^{\otimes \mathbb{N}}\) with the norm
\[
\|\alpha\| := \sum_{k \in \mathcal{K}} \alpha_k
\]
and the partial order
\[
\alpha \leq \beta \text{ if and only if } \forall k \in \mathcal{K}. \alpha_k \leq \beta_k.
\]
Multi-indices are multisets with extra structure.
**Definition 570.** Fix a commutative semiring \( R \) and a set \( \mathcal{X} \) of symbols, which we will call indeterminates.

Let \( M \) be the free \( R \)-semimodule over \( \mathcal{X} \), written multiplicatively. We will call the members of \( M \) monomials. Using a multi-index \( \gamma \) over \( \mathcal{X} \), every monomial can be written as

\[
\prod_{X \in \mathcal{X}} X^{\gamma_X},
\]

where \( \gamma_X \) are the coefficients in \( R^{\oplus \mathcal{X}} \) of the monomial.

The polynomial algebra or polynomial semiring \( R[\mathcal{X}] \) for the given indeterminates is the free \( R \)-semimodule over \( M \). That is, a polynomial \( p \in R[\mathcal{X}] \) is an \( R \)-linear combination of monomials, and we denote polynomials by

\[
p(\mathcal{X}) = \sum_{\gamma} a_{\gamma} \prod_{X \in \mathcal{X}} X^{\gamma_X}, \tag{203}
\]

We call \( a_{\gamma} \) the coefficients of the polynomial. We use the components of the multi-index as powers in the monomials, but we use \( \gamma \) itself as an index for the coefficient \( a_{\gamma} \). Unfortunately, multi-indices are sometimes confusing, but often their brevity outweighs the possible confusion.

We do not ignore the structure of \( M \). We conflate exponentiation in \( M \) in the sense of definition 447(e) with exponentiation in \( R[\mathcal{X}] \) in the sense of definition 525(e). Multiplication in \( M \) motivates us to define multiplication in \( R[\mathcal{X}] \) via a convolution of the coefficients. We define the product of \( p(\mathcal{X}) \) from (203) with

\[
q(\mathcal{X}) = \sum_{\gamma} b_{\gamma} \prod_{X \in \mathcal{X}} X^{\gamma_X} \tag{204}
\]

as

\[
[pq](\mathcal{X}) := \sum_{\gamma} \left( \sum_{\delta+\eta=\gamma} a_{\delta} b_{\eta} \right) \prod_{X \in \mathcal{X}} X^{\gamma_X}. \tag{205}
\]

We simultaneously use multi-indices as vectors with pointwise summation (i.e. \( \delta + \eta = \gamma \)) and as indices of coefficients (i.e. \( a_{\delta} \) and \( b_{\eta} \)).

We avoid writing the embedding \( \iota : \mathcal{X} \to R[\mathcal{X}] \), but it is sometimes beneficial to denote it explicitly, for example in theorem 571 (Polynomial algebra universal property).

**Theorem 571 (Polynomial algebra universal property).** Fix a commutative semiring \( R \) and a set \( \mathcal{X} \) of indeterminates. The polynomial algebra \( R[\mathcal{X}] \) is the unique up to a unique isomorphism commutative algebra that satisfies the following universal mapping property:

For every commutative \( R \)-algebra \( M \) and every function \( e : \mathcal{X} \to M \), there exists a unique \( R \)-algebra homomorphism \( \Phi_e : R[\mathcal{X}] \to M \) such that the following diagram commutes:

\[
\begin{array}{ccc}
\mathcal{X} & \xrightarrow{e} & M \\
\downarrow \iota & & \\
R[\mathcal{X}] & \xrightarrow{\Phi_e} & \end{array}
\]

(206)
The function \( e \) evaluates each indeterminate in \( M \), while \( \Phi_e \) substitutes this value in every polynomial. We call \( \Phi_e \) the substitution homomorphism corresponding to the variable assignment \( e \). We can parameterize this by the evaluation functions to obtain the functional evaluation homomorphism

\[
\Phi : R[X] \to \text{fun}(M^X, M)
\]

\[
\Phi(p) := (e \mapsto \Phi_e(p))
\]

We call the values of \( \Phi \) polynomial functions. Given elements \( x_1, \ldots, x_n \) of \( M \), we write

\[
p(x_1, \ldots, x_n)
\]

rather than

\[
\Phi(p)(x_1, \ldots, x_n).
\]

Via remark 1198, \( R[-] \) becomes left adjoint to the forgetful functor

\[
U : \text{CAlg}_R \to \text{Set}.
\]

The action of \( R[-] \) on morphisms is given by \( \Phi \).

Proof. For every indeterminate \( X \), we want

\[
\Phi_e(\epsilon(X)) = f(X).
\]

This suggests defining \( \Phi_e \) for the polynomial

\[
p(X) = \sum_y a_y \prod_{X \in X} t(X)^{y_X}
\]

as the evaluation

\[
\Phi_e(p) := \sum_y a_y \prod_{X \in X} f(X)^{y_X}.
\]

We discuss well-definedness of infinitary operations in direct sums in remark 794 (c iii). □

Remark 572. As we saw in definition 570 and theorem 571 (Polynomial algebra universal property), there is no formal problem in defining polynomial algebras over infinitely many indeterminates.

There is a problem, however. Polynomials in one indeterminate, which we will call univariate in accordance to definition 966 (c), have a well-ordering on their monomials, induced by the degree of their monomials. This is defined and discussed in section 10.7 (Univariate polynomials).

Polynomials in more than one variable do not have a well-ordering by default. If the indeterminates themselves are well-ordered, as is the case for finitely many indeterminates, we may introduce, for example, a reverse lexicographic order on the monomials. Furthermore, for finitely many variables, proposition 573 (b) allows us to use natural number induction on the number of variables in order to prove statements about multivariate polynomial rings.

For infinitely many, especially uncountably many variables, however, the theory is seriously crippled by the lack of the tools described above. For this reason, only polynomials in finitely many variables are often considered.
Proposition 573. The following are basic properties of polynomial semirings:

(a) If \( \mathcal{X} \) is empty, \( R[\mathcal{X}] \cong R \).

(b) The polynomial algebras \( R[\mathcal{X}] \) and \( R[\mathcal{X} \setminus \{X_0\}][X_0] \) are isomorphic for any \( X_0 \in \mathcal{X} \) (in case \( \mathcal{X} \) has more than one member).

In particular,
\[
R[X_1, \ldots, X_{n-1}][X_n] \cong R[X_1, \ldots, X_n].
\]

(c) The univariate polynomial semiring \( R[X] \) is entire if and only if \( R \) is entire.

(d) If \( R \) is entire, the units in \( R[X_1, \ldots, X_n] \) are precisely the (embeddings of) the units of \( R \).

Proof.

Proof of 573 (a). Trivial.

Proof of 573 (b). Polynomials in \( R[\mathcal{X}] \) have the form
\[
p(\mathcal{X}) = \sum_{k=0}^{\infty} \sum_{\gamma} \left( \sum_{X \in \mathcal{X} \setminus \{X_0\}} a_{(k,\gamma)} X^{r_X} X_0^k \right) X_0^k,
\]
where \( \gamma \) is a multi-index on \( \mathcal{X} \).

Due to associativity, commutativity and distributivity, this can be rewritten as
\[
p(\mathcal{X}) = \sum_{k=0}^{\infty} \left( \sum_{\gamma} a_{(k,\gamma)} \prod_{X \in \mathcal{X} \setminus X_0} X^{r_X} \right) X_0^k.
\]

This shows how \( R[\mathcal{X}] \) can be embedded into \( R[\mathcal{X} \setminus \{X_0\}][X_0] \). This embedding is surjective because the coefficients \( a_\gamma \) range through \( R \). Therefore, the embedding is an isomorphism.

Proof of 573 (c).

Proof of sufficiency. Since \( R \) is an \( R \)-subalgebra of \( R[X] \), if the latter is entire, so is the former.

Proof of necessity. Suppose that \( R \) is entire and that \( R[X] \) isn’t. Then there exist nonzero polynomials \( p(X) \) and \( q(X) \) such that \( p(X)q(X) = 0 \). If \( a_n \) is the leading coefficient of \( p(X) \) and \( b_m \) — of \( q(X) \), the leading coefficient of \( p(X)q(X) \) is \( a_nb_m \). Since \( p(X)q(X) \) is the zero polynomial, \( a_nb_m = 0 \), which contradicts the assumption that \( R \) is entire.

Therefore, \( R[X] \) is entire.

Proof of 573 (d). As in proposition 573 (c), it is sufficient to prove the statement for one indeterminate.

Clearly every constant is invertible as a constant polynomial.

Now suppose that \( p(X)q(X) = 1 \). By definition of multiplication, the product has only one nonzero coefficient. Since \( R \) is entire, it follows that both \( p(X) \) and \( q(X) \) have only one nonzero coefficient, and are hence constants. \( \Box \)
Example 574. We list several examples of polynomials over semirings.

(a) Consider the polynomial \( p(X) := aX^2 + bX + c \) in \( \mathbb{N}[X] \). A function from the set \{X\} to \( \mathbb{N} \) corresponds to an element of \( \mathbb{N} \), and hence evaluating the polynomial is done by simply replacing \( X \) symbolically in \( p \) and then evaluating the obtained syntax tree.

We seek the roots of \( p(X) \). We will only formally define roots in definition 641; for the purposes of the example, a root is a natural number \( n \) such that \( \Phi_n(p) = 0_R \).

By theorem 41 (Fundamental theorem of algebra) and definition 707 (e), \( p \) has two roots in the complex plane. That is, we regard \( \mathbb{C} \) as an algebra over \( \mathbb{N} \) and use theorem 571 (Polynomial algebra universal property) to obtain a polynomial function on \( \mathbb{C} \). Furthermore, over the complex numbers the roots can be explicitly found using

\[
-\frac{b \pm \sqrt{b^2 - 4ac}}{2a}.
\]

Finding a root of \( p \) over the natural numbers cannot be done in general, however. If \( p(n) = 0 \), by the ordering of the natural numbers we have

\[
p(n) = an^2 + bn + c \geq c,
\]

and hence \( c \) must necessarily be 0. If \( c = 0 \), then zero is a root of the polynomial \( p(X) = aX^2 + bX \).

Now let \( n \) be any root of \( p \). We have

\[
an^2 + bn \geq bn,
\]

and hence \( bn \) must also be 0. Thus, either \( b = 0 \) or \( n = 0 \). If we want a root other than \( n \), both \( a \) and \( b \) must be 0.

Therefore, the only natural number solution to the quadratic equation is 0, and it is only a solution if \( c = 0 \).

(b) Consider again the polynomial \( p(X) := aX^2 + bX + c \) over \( \mathbb{N} \), but this time evaluated over the min-plus semiring \((\mathbb{N} \cup \{\infty\}, \min, +)\).

Expressed via the standard natural number operations, this polynomial becomes

\[
\min\{2X + a, X + b, c\}.
\]

This allows us to express certain optimization problems via polynomials.

This polynomial has a root if and only if \( a = b = c \). Roots in the tropical semiring are not very interesting, however.

Proposition 575. For a set \( A \) in an \( R \)-algebra \( M \), the generated subalgebra of \( A \), defined as the \( R \)-subalgebra generated by \( A \) in the sense of definition 852, equals the set

\[
\bigcup\{R[a_1, \ldots, a_n] \mid a_1, \ldots, a_n \in A\}
\]
obtained by evaluating all multivariate polynomials over $R$ with elements of $A$.

The $\mathbb{N}$-subalgebras of $M$ correspond to sub-semirings and the $M$-subalgebras correspond to ideals.

Compare this result to proposition 554 for modules.

**Proof.** Similar to proposition 554.

**Proposition 576.** Let $R \subseteq S$ be commutative semirings and let $A \subseteq S$ be an arbitrary subset.

Fix a set $\mathcal{X}$ of indeterminates and a bijective function $e : \mathcal{X} \to A$ and consider the evaluation homomorphism

$$
\Phi_e : R[\mathcal{X}] \to S.
$$

The image $R[A]$ of $\Phi_e$ is the smallest super-semiring of $R$ that contains $A$.

We say that $R[A]$ is obtained by adjoining the elements of $A$ to $R$.

**Proof.** Follows from proposition 575.

**Example 577.** Continuing example 574 (a), consider the polynomial equation

$$X + 1 = 0.$$

It has no natural number root as a consequence of (PA2).

It does have an integer root, however, $-1$. We can adjoin $-1$ to the semiring $\mathbb{N}$ to obtain the semiring $\mathbb{N}[-1]$. But this latter semiring is (isomorphic to) $\mathbb{Z}$.

Therefore, $\mathbb{Z}$ is the smallest extension of $\mathbb{N}$ that contains a root to the polynomial $X + 1$.

This example extends to the theory of transcendental and algebraic elements of fields.

**Definition 578.** If we extend the concept of polynomials to allow countably many nonzero terms, we obtain a set $R[[\mathcal{X}]]$ which we call the formal power series over $R$ with indeterminates from the set $\mathcal{X}$.

The evaluation homomorphism defined in theorem 571 (Polynomial algebra universal property) is problematic, however, since algebraic operations are finitary by nature. This is discussed in remark 794 (c), along with how sometimes we can make sense of infinitary algebraic operations.
10.5. Rings

**Definition 579.** A ring is a semiring with additive inverses. More precisely, this means that the additive monoid is a group.

As for semirings, rings can also be nonunital, with ring ideals being the main example. Rings have the following metamathematical properties:

(a) We can construct a first-order theory for rings by adding a unary functional symbol – and the involution axiom (167) to the theory of semirings.

(b) A first-order homomorphism between the rings $R$ and $T$ is a semiring homomorphism $\varphi : R \to T$ that additionally preserves additive inverses.

As shown in proposition 458, this condition is not only redundant, but the structure of a ring rather than semiring also automatically implies that $\varphi(0_R) = 0_T$.

(c) The set $A \subseteq R$ is a submodel of $R$ if it is both a sub-semiring of $R$ and an additive submonoid of $R$.

As a consequence of proposition 863, the image of a ring homomorphism is a subring of its range.

(d) The trivial ring is the trivial pointed set $\{0\}$.

(e) If multiplication is commutative, we call the ring itself commutative. Unless multiplication corresponds to function composition, most rings we will encounter will be commutative.

(f) The corresponding category of $\mathcal{U}$-small models $\mathcal{U}$-Ring is concrete over $\mathcal{U}$-SRing. We denote the category of commutative rings by $\textbf{CRing}$.

Unlike the category $\textbf{Grp}$ of groups, $\textbf{Ring}$ is not as well-behaved. Nevertheless, kernels and quotients of rings are commonly established concepts.

The category of unital rings does not have a zero object, but the category of nonunital rings does, and we will sometimes consider nonunital ring homomorphisms between unital rings. That is, ring homomorphisms that may not preserve the multiplicative identity.

(g) The kernel of a ring homomorphism $\varphi : R \to T$ is simply its zero locus $\varphi^{-1}(0_T)$. This is precisely the kernel of the additive group in the sense of definition 455 (h), and the categorical kernel in the category of nonunital rings.

Furthermore, $\ker \varphi$ is an ideal of $R$ because, if $x \in \ker \varphi$,

$$\varphi(xy) = \varphi(x)\varphi(y) = 0_S \varphi(y) = 0_S,$$

and thus $xy \in \ker \varphi$.

Despite being categorical kernels only in the category of nonunital rings, the kernel is defined and used mainly for unital ring homomorphism.
(h) The **categorical cokernel** of homomorphism $\varphi : R \rightarrow T$ in the category of nonunital rings is, similarly to the case for groups in definition 455 (i), a partition of $T$ induced by the image of $\varphi$.

This is not merely the cokernel $T/\text{img } \varphi$ of the additive group, however. Multiplication induces an additional restriction on congruences: $x \equiv x'$ and $y \equiv y'$ together imply $xy \equiv x'y'$. Hence, $[x][y] = [xy]$. Denote the coset $[0_s]$ by $I$. We have $I[x] = [0x] = I$, therefore the cokernel inherits absorption from $T$.

Additive subgroups of $T$ that absorb multiplication are precisely the **two-sided ideals** of $T$. Hence, $I$ is the ideal generated by $\text{img } \varphi$. From the general case for groups it follows that quotient ring cosets have the form $x + I$.

Finally, given an ideal $I$ of an arbitrary ring $R$, we can define the **quotient ring** $R/I$ as the cokernel of the inclusion $i : I \rightarrow R$. That is, $R/I$ consists of the cosets $x + I$ for $x \in R$. In practice, quotients are conveniently characterized by theorem 649 (Quotient algebra universal property).

Somewhat similarly to proposition 457 (h) for groups, the kernel $\ker \pi$ of the canonical projection $\pi(x) := x + I$ is the ideal $I$ itself.

Despite being categorical cokernel only in the category of nonunital rings, the quotient $R/I$ is defined and used mainly for unital rings.

Fortunately, for unital ring $R$, the quotient $R/I$ is also unital. The projection morphism is an epimorphism by proposition 1216, and hence $R/I$ is a categorical quotient object.

(i) Analogously to simple groups, if the only proper ideal of $R$ is the trivial ideal $\{0_R\}$, we say that $R$ is a **simple ring**.

The trivial ring itself is not simple, because it has no proper ideals.

**Proposition 580.** Given an ideal $I$ in a ring $R$, we have $x + I = y + I$ if and only if $x - y \in I$.

**Proof.** Trivial.

**Proposition 581.** For a positive integer $n > 1$, we extend the group $\mathbb{Z}_n$ of integers modulo $n$ with the operation

$$x \circ y := \text{rem}(xy, n).$$

Then $\mathbb{Z}_n$ is a commutative ring called the **ring of integers modulo** $n$.

**Proof.** Note that

\[
\begin{align*}
\text{rem}(x, n) \text{rem}(y, n) & \equiv (\text{mod } n) \\
\equiv (x - n \text{quot}(x, n))(y - n \text{quot}(y, n)) & \equiv (\text{mod } n) \\
\equiv xy - n \text{quot}(x, n) - n \text{quot}(y, n) + n^2 \text{quot}(x, n) \text{quot}(y, n) & \equiv (\text{mod } n) \\
\equiv xy.
\end{align*}
\]

The proof that multiplication in $\mathbb{Z}_n$ is associative, unital and commutative becomes trivial.
We will prove that multiplication distributes over addition. Fix \(x, y, z \in \mathbb{Z}_n\). We have
\[
(x \oplus y) \odot z = \text{rem}((x \oplus y)z, n) = \text{rem}(\text{rem}(x + y, n)z, n) = \text{rem}((x + y - n \text{quot}(x + y, n))z, n) = \text{rem}((x + y)z, n).
\]

and
\[
(x \odot z) \oplus (y \odot z) = \text{rem}([(x \odot z) + (y \odot z)], n) = \text{rem}([xz - n \text{quot}(xz, n) + yz - n \text{quot}(yz, n)], n) = \text{rem}(xz + yz, n) = \text{rem}((x + y)z, n).
\]

Hence,
\[
(x \oplus y) \odot z = (x \odot z) \oplus (y \odot z).
\]

\[\square\]

**Proposition 582.** Similarly to how \(\mathbb{N}\) is an initial object in the category \(\text{Sring}\) of semirings, \(\mathbb{Z}\) is an initial object in the category \(\text{Ring}\) of rings.

**Proof.** Follows from proposition 530 with the addition of \((-n)x = -nx\).

**Definition 583.** We define the characteristic \(\text{char}(R)\) of a ring \(R\) via the following equivalent definitions:

(a) \(\text{char}(R)\) is the unique nonnegative integer \(n\) for which \(\mathbb{Z}_n\) can be embedded into \(R\). That is,
\[
\mathbb{Z}_n \cong \mathbb{Z} / \ker \iota,
\]
where \(\iota\) is the homomorphism from the integers defined via (??).

We use here that \(\mathbb{Z}_0\) is the trivial ring.

(b) \(\text{char}(R)\) is the order of the additive group of \(R\) and, if it exists, and 0 otherwise. That is, \(\text{char}(R)\) is the smallest positive integer \(n\) such that \(n \cdot 1_R = 0_R\) and \(\text{char}(R) = 0\) if \(0_R\) cannot be obtained in this way.

**Proof of correctness.**

**Proof of equivalence of ?? and 583 (b).** Let \(n\) be such that
\[
n\mathbb{Z} = \ker \iota.
\]

In particular, \(\iota(0) = \iota(n)\).

If \(n = 0\), \(\ker \iota\) is a trivial group and \(\iota\) is an embedding. Then there cannot exist a positive integer \(n\) such that
\[
n \cdot 1_R = 0_R.
\]

286
Otherwise, $n$ is the smallest positive integer such that
\[ n \cdot 1_R = 0 \cdot 1_R = 0_R. \]

\[ \square \]

**Proposition 584.** If $\varphi : R \to S$ is a ring embedding, then $S$ inherits its characteristics from $R$. 

**Proof.** First suppose that $R$ has positive characteristic $n$. Then $n \cdot 1_R = 0_R$, which implies $n \cdot \varphi(1_R) = \varphi(0_R)$, hence $\text{char}(T) \leq n$. But $\varphi$ is an embedding, hence if $k \cdot 1_R \neq 0_R$, then $k \cdot \varphi(1_R) \neq \varphi(0_R)$. 

This implies that $\text{char}(S) \geq \text{char}(R)$, which in turn shows that $\text{char}(S) = \text{char}(R)$. 

If $R$ has characteristic zero, then $\varphi : \mathbb{N} \to R$ is an embedding and thus $\varphi \circ \iota : \mathbb{N} \to S$ is also an embedding. It is unique as shown in proposition 582. Therefore, $S$ also has characteristic zero. 

\[ \square \]

**Example 585.** The following are examples of ring characteristics:

(a) The integers $\mathbb{Z}$ have characteristic $\text{char}(\mathbb{Z}) = 0$ because $\iota$ is an isomorphism. Consequently, by proposition 584, any superring of $\mathbb{Z}$ has characteristic zero, most notably the fields $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. 

(b) The ring $\mathbb{Z}_n$ of integers modulo $n$ has characteristic $\text{char}(\mathbb{Z}_n) = n$ because of proposition 517. 

(c) An algebra $M$ over a nontrivial commutative unital ring $R$ has the same characteristic as $R$ because of the canonical embedding of $R$ in $M$. In particular, the polynomial ring $R[X]$ has the same characteristic as their ring.

**Proposition 586.** The Grothendieck completion $\overline{R}$ of the additive monoid of a semiring $R$ becomes a ring with the operation
\[ [(a, b)] \odot [(c, d)] := [(ac + bd, ad + bc)]. \]

This definition is motivated in the proof of theorem 587 (Grothendieck semiring completion universal property). 

**Proof.** Multiplication on $R$ does not depend on the representative of the equivalence class. Indeed, let $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$. Then there exist $u$ and $v$ such that
\[ a + b' + u = a' + b + u, \]
\[ c + d' + v = c' + d + v. \]

Then
\[
\begin{align*}
ac + b'c + uc + a'd + bd + ud + a'c + a'd' + a'v + b'c' + b'd + b'v &= \\
= (a + b' + u)c + (a' + b + u)d + a'(c + d' + v) + b'(c' + d + v) &= \\
= (a' + b + u)c + (a + b' + u)d + a'(c' + d + v) + b'(c + d' + v) &= \\
= a'c + bc + uc + ad + b'd + ud + a'c' + a'd + a'v + b'c + b'd' + b'v.
\end{align*}
\]
Therefore,

\[(a \cdot c + b \cdot d, a \cdot d + b \cdot c) \sim (a' \cdot c' + b' \cdot d', a' \cdot d' + b' \cdot c').\]

Associativity and distributivity in \(\overline{R}\) are inherited from \(R\).

**Theorem 587** (Grothendieck semiring completion universal property). The Grothendieck completion \(\overline{R}\) of a semiring \(R\) satisfies the following universal mapping property:

For every ring \(T\) and every semiring homomorphism \(\varphi : R \rightarrow T\), there exists a unique ring homomorphism \(\overline{\varphi} : \overline{R} \rightarrow T\) such that the following diagram commutes:

\[
\begin{array}{ccc}
R & \xrightarrow{\varphi} & S \\
\downarrow & & \downarrow \\
\overline{R} & \xrightarrow{\overline{\varphi}} & T \\
\end{array}
\]

(207)

Via remark 1198, \(\overline{\ }\) becomes left adjoint to the forgetful functor

\(U : CRing \rightarrow CSRing\).

Compare this result to theorem 521 (Grothendieck monoid completion universal property).

**Proof.** Theorem 521 (Grothendieck monoid completion universal property) suggests the definition

\[\overline{\varphi}([(a, b)]) := \varphi(a) - \varphi(b).\]

We must only show that \(\overline{\varphi}\) is a ring homomorphism. Clearly

\[\overline{\varphi}([(1, 0)]) = \varphi(1) - \varphi(0),\]

which implies that \(\overline{\varphi}\) preserves multiplicative identities. Also,

\[
\overline{\varphi}([(a, b)] \odot [(c, d)]) = \overline{\varphi}([(a \cdot b + c \cdot d, a \cdot d + b \cdot c)]) =
\]

\[= \varphi(a \cdot b + c \cdot d) - \varphi(a \cdot d + b \cdot c) =
\]

\[= \varphi(c)(\varphi(d) - \varphi(b)) - \varphi(a)(\varphi(d) - \varphi(b)) =
\]

\[= (\varphi(c) - \varphi(a))(\varphi(d) - \varphi(b)) =
\]

\[= \overline{\varphi}([(a, c)])\overline{\varphi}([(b, d)]).\]

\[\square\]

**Proposition 588.** The Grothendieck completion \(\overline{R}\) of a semiring \(R\) satisfies the following basic properties:

(a) If \(R\) is commutative, so is \(\overline{R}\).
(b) If \( R \) is entire, so is \( \overline{R} \).

Proof.

**Proof of 588 (a).** This is clear from the definition of multiplication.

**Proof of 588 (b).** Suppose that

\[
[(a, b)] \cdot [(c, d)] = [(0, 0)]
\]

Then there exists an element \( u \) in \( R \) such that

\[
(ac + bd) + 0 + u = 0 + (ad + bc) + u.
\]

Suppose that \( d = c + e \). Then

\[
ac + b(c + e) = a(c + e) + bc
\]

and

\[
(ac + bc) + be = (ac + bc) + ae.
\]

Cancelling \( e \), we obtain that \( a = b \). But \([a, b] = [(0, 0)]\). \( \square \)

**Definition 589.** Let \( R \) be an arbitrary ring. We define the **commutator** of the elements \( x \) and \( y \) as

\[
[x, y] := xy - yx.
\]

The **commutator ideal** \([R, R]\) of \( R \) is the two-sided ideal generated by all the commutators in \( G \).

**Theorem 590 (Ring abelianization universal property).** The quotient \( R/[R, R] \) of a ring \( R \) by its commutator ideal \([R, R]\) is a commutative ring, which we call the **abelianization** of \( R \), and satisfies the following universal mapping property:

For every commutative ring \( T \) and every ring homomorphism \( \varphi : R \to T \), \( \varphi \) uniquely factors through \( R/[R, R] \). That is, there exists a unique ring homomorphism \( \overline{\varphi} : R/[R, R] \to T \) such that the following diagram commutes:

\[
\begin{array}{ccc}
R & \xrightarrow{\varphi} & S \\
\downarrow{\pi_R} & & \downarrow{\overline{\varphi}} \\
R/[R, R] & &
\end{array}
\]

(208)

Via remark 1198, the abelianization functor becomes left adjoint to the forgetful functor

\( U : \text{CRing} \to \text{Ring} \).

Compare this result to theorem 523 (Group abelianization universal property).
Proof. This is a refinement of theorem 523 (Group abelianization universal property), and we only need to show that $R/[R,R]$ is a commutative ring. For $x$ and $y$ in $R$, since $yx - xy \in I$, we have

$$(x + I)(y + I) = (xy + I) = (yx + xy - xy + I) = (yx + I) = (y + I)(x + I).$$

\[\square\]

**Definition 591.** We call the subset of the semiring $R$ a **multiplicative set** if it contains $1_R$ and, furthermore, it is closed under multiplication.

**Proposition 592.** The ideal $P$ in the commutative semiring $R$ is **prime** if and only if $R \setminus P$ is a multiplicative set.

Not all multiplicative sets are obtained as complements of prime ideals — see example 595 (b).

Proof. By proposition 558 (a), $P$ is a proper ideal if and only if $1_R \in R \setminus P$.

By proposition 558 (h), $P$ is prime if and only if $x, y \in R \setminus P$ implies $xy \in R \setminus P$. \[\square\]

**Definition 593.** Let $R$ be a commutative ring and let $S \subseteq R$ be a multiplicative set.

Define the equivalence relation $(r, s) \sim (r', s')$ on $R \times S$ to hold if and only if there exists some $u \in S$ such that $urs' = ur's$.

Consider the set

$$S^{-1}R := R \times S / \sim,$$

whose cosets we will denote by $r/s$ rather than $[(r, s)]$.

Define on $S^{-1}R$ the operations

$$\begin{align*}
\frac{a}{b} + \frac{c}{d} &= \frac{ad + bc}{bd}, \\
\frac{a}{b} \cdot \frac{c}{d} &= \frac{ac}{bd}, \\
\end{align*}$$

and the canonical inclusion

$$\iota : R \rightarrow S^{-1}R$$

$$\iota(r) := \frac{r}{1_R}.$$ 

This ring is called the **localization** of $R$ with respect to $A$; we denote it by $S^{-1}R$. In case $S$ is the complement of a **prime ideal**, we may denote the localization by $R_P$ (or $R_p$ if $P = \langle p \rangle$).

The image under $\iota$ of every element $s$ of $S$ is invertible in $S^{-1}R$, and we call the inverse $1_R/s$ the **reciprocal** of $s$.

This construction is very similar to the **Grothendieck completion** of a monoid or semiring, although with notable differences — the set $S$ may be a strict subset of $R$, and addition in the Grothendieck completion corresponds to multiplication in the localization, while addition in the completion has no analogy.
Proof of correctness. The proof that $\sim$ is an equivalence relation is the same as in definition 520. The result is then a ring if the operations are well-defined.

We will show that both operations are well-defined. Let $uab' = u'ab$, meaning that $(a, b) \sim (a', b')$ and hence $a/b = a'/b'$, and let $vcd' = v'c'd$.

For addition, we have

$$uv(ad + bc)b'd' = vdd'(uab') + ubb'(vcd') = vdd'(ua'b) + ubb'(vc'd) = uv(a'd' + b'c')bd,$$

hence $(ad + bc, bd) \sim (a'd' + b'c', b'd')$.

The proof for correctness of multiplication is the same as the proof of correctness of addition in definition 520.

\[\text{Theorem 594 (Ring localization universal property).} \text{ The localization of } R \text{ by } S \text{ satisfies the following universal mapping property:}\]

For every commutative ring $T$ and every ring homomorphism $\varphi : R \to T$ such that $\varphi(s)$ is invertible in $T$ for every $s \in S$, $\varphi$ uniquely factors through $S^{-1}R$. That is, there exists a unique ring homomorphism $\tilde{\varphi} : S^{-1}R \to T$ such that the following diagram commutes:

$$\begin{array}{ccc}
R & \xrightarrow{\varphi} & T \\
\downarrow & & \downarrow \\
S^{-1}R & \xrightarrow{\tilde{\varphi}} & T
\end{array}$$  \hspace{1cm} (209)

Proof. The condition suggests the definition

$$\tilde{\varphi}\left(\frac{r}{s}\right) := \varphi(r)\varphi(s)^{-1}.$$  \hspace{1cm} \[\square\]

Example 595. We list several examples of commutative ring localization.

(a) If $S$ contains $0_R$, then $S^{-1}R$ is the trivial ring.

(b) The localization $S^{-1}\mathbb{Z}$ by the set $S := \{2^n \mid n \geq 0\}$ is (a ring isomorphic to) the rational numbers with denominators that are powers of two. This is an example of a multiplicative set that is not the complement of a prime ideal.

This ring is isomorphic to the ring $\mathbb{Z}[1/2]$ obtained by adjoining the rational number $1/2$ to $\mathbb{Z}$.

(c) Let $p$ be a prime number. The localization $S^{-1}\mathbb{Z}$ by $S := \mathbb{Z} \setminus \langle p \rangle$ is (a ring isomorphic to) the rational numbers with denominators coprime to $p$.

For $p = 2$, this localization consists of rational numbers whose denominator is an odd number.
Proposition 596. Ring localization has the following basic properties:

(a) Localization preserves ideals. More precisely, given a commutative ring $R$, a multiplicative set $S$ and an ideal $I$, the set

$$S^{-1}I := \left\{ \frac{r}{s} \middle| r \in I \text{ and } s \in S \right\}$$

is an ideal of the localization $S^{-1}R$.

(b) The map $I \mapsto S^{-1}I$ is a strict order isomorphism between the set of prime ideals of $R$ not intersecting $S$ and the set of all prime ideals of $S^{-1}R$.

(c) The localization $R_P$ by a prime ideal $P$ has a unique maximal ideal $S^{-1}P$ (here $S := R \setminus P$).

(d) The canonical inclusion $\iota : R \to S^{-1}R$ is injective if and only if $S$ contains no zero divisors.

Proof.

Proof of 596 (a). Trivial since $S$ is closed under multiplication.

Proof of 596 (b). Let $P$ be a prime ideal in $R$ disjoint from $S$. By proposition 596 (a), $S^{-1}P$ is an ideal of $S^{-1}R$. If the product $ac/bd$ belong to $S^{-1}P$, then $ac \in P$ and $bd \in S$. Since $P$ is prime, $a \in P$ or $c \in P$. If $a \in P$, then $ba \in P$ and $a/d = ba/bd \in S^{-1}P$; if $c \in P$, we proceed analogously. Thus, $S^{-1}P$ is a prime ideal, i.e. the image under $I \mapsto S^{-1}I$ of a prime ideal is a prime ideal.

Proof of injectivity. Let $S^{-1}P = S^{-1}Q$ for prime ideals $P$ and $Q$ disjoint from $S$. Suppose that $P \setminus Q$ contains at least one element, say $p$. Then $\iota(p) = p/1$ belongs to both $S^{-1}P$ and $S^{-1}Q$; hence, $Q$ contains an element $q$ such that, for some $s \in S$ and $u \in S$,

$$p \cdot s \cdot u = 1 \cdot q \cdot u.$$

Since $Q$ is an ideal, $qu \in Q$, and hence $psu \in Q$. But neither $p$, $s$, nor $u$ belong to $Q$, which contradicts the assumption that $Q$ is prime. Therefore, $P \setminus Q$ is empty. Generalizing, we obtain that $I \mapsto S^{-1}I$ is injective on prime ideals.

Proof of surjectivity. Fix a prime ideal $T$ in $S^{-1}R$ and let $P$ be the set of numerators in $T$, i.e. if $p/s \in T$, then $p \in P$. We will show that $P$ is a prime ideal; clearly $T = S^{-1}P$.

Clearly $0_R \in P$. Let $a, c \in P$. Then there exist $b, d \in S$ such that $a/b$ and $c/d$ belong to $T$. But $T$ is closed under multiplication with members of $R$, hence $a/1_R = b(a/b)$ and $c/1_R = d(c/d)$ also belong to $T$. Then their sum $(a + c)/1$ belongs to $T$, and hence also to $P$. Thus, $P$ is closed under addition. We analogously obtain that it is closed under multiplication.

We have shown that $P$ is an ideal in $R$. We must show that it is a prime ideal. Let $ac \in P$. Then

$$\frac{a}{b} \cdot \frac{c}{d} \in T$$

for some $b, d \in S$. Hence, $a/b$ or $c/d$ belongs to $T$, implying that $a \in P$ or $c \in P$.

Proof of monotonicity. Follows from proposition 1227.
**Proof of 596 (c).** In the localization $R_P$ be a prime ideal, all members of $P$ become invertible. Hence, a maximal ideal cannot contain members of $P$. By proposition 596 (a), $S^{-1}P$ is an ideal, therefore it must be the largest proper ideal.

**Proof of 596 (d).** Let $sr = 0$ for $s \in S$. Then $\iota(s) = s/1_R$ is invertible in $S^{-1}R$ and hence

$$
\frac{0_R}{1_R} = \frac{sr}{1_R} = \frac{1_R}{s} \cdot \frac{sr}{1_R} = \frac{r}{1_R}.
$$

Hence, $\iota(r) = \iota(0_R)$.

It follows that $\iota$ is injective if and only if $S$ contains no zero divisors.

**Definition 597.** If every nonzero element of a ring is invertible, we call it a division ring.

**Proposition 598.** A nontrivial division ring is entire.

**Proof.** Let $xy = 0$. If $x$ is nonzero, multiplying both sides by $x^{-1}$, we obtain $y = 0$. Analogously, $y \neq 0$ implies that $x = 0$. In all cases, either $x$ or $y$ is necessarily zero.

Therefore, the ring has no nontrivial zero divisors.

**Definition 599.** We will call the nontrivial ring $\mathbb{K}$ a field if any of the following equivalent conditions hold:

(a) $\mathbb{K}$ is commutative and simple.

(b) $\mathbb{K}$ is a commutative division ring.

Fields have the following metamathematical properties:

(a) We can construct a first-order theory for fields by adding to the theory of rings the axioms $\neg(0 \doteq 1)$ and

$$(\xi \doteq 0) \lor \exists \eta. (\xi \cdot \eta \doteq 1).$$

These axioms are not positive formulas, hence fields automatically get worse metamathematical properties than rings, for example.

(b) A first-order homomorphism between fields is simply a unital ring homomorphism.

(c) If for two fields $\mathbb{k}$ and $\mathbb{K}$ are have $\mathbb{k} \subseteq \mathbb{K}$, we say that $\mathbb{K}$ is a field extension of $\mathbb{k}$ and that $\mathbb{k}$ is a subfield of $\mathbb{K}$. In particular, if $\mathbb{K} = \mathbb{k}$, we say that the extension is trivial.

(d) The category of $\mathcal{U}$-small fields $\mathcal{U}$-$\text{Field}$ is a full subcategory of $\mathcal{U}$-$\text{CRing}$ with objects restricted to fields.

**Proof of correctness.** The equivalence of definitions follows from proposition 558 (b).

**Proposition 600.** Let $D$ be an integral domain. The localization of $D$ at the zero ideal $\{0_R\}$ is a field, which we call the field of fractions of $D$. 

293
Proof. By ??, the localization by the prime ideal \( \{0_R\} \) has whose only maximal ideal is \( S^{-1}\{0_R\} \). Since \( 0_R \) is absorbing, \( S^{-1}\{0_R\} \) is again the zero ideal. Therefore, it is the only proper ideal of the localization \( S^{-1}D \), and hence the localization is a simple ring.

Since \( D \) is an integral domain, by proposition 596 (d), \( S^{-1}D \) is a superring of \( D \). It is therefore a nontrivial commutative simple ring, and thus it satisfies definition 599 (a). \( \square \)

**Theorem 601** (Field of fractions universal property). The field of fractions \( \mathbb{K} \) of the integral domain \( D \) satisfies the following universal mapping property:

For every field \( \mathbb{L} \) and every ring homomorphism \( \varphi : D \rightarrow \mathbb{L} \), \( \varphi \) uniquely factors through \( \mathbb{K} \). That is, there exists a unique field homomorphism \( \tilde{\varphi} : \mathbb{K} \rightarrow \mathbb{L} \) such that the following diagram commutes:

\[
\begin{array}{ccc}
R & \xrightarrow{\varphi} & \mathbb{L} \\
\downarrow{id} & & \downarrow{\tilde{\varphi}} \\
\mathbb{K} & & \\
\end{array}
\]

(211)

Proof. This is simply a special case of theorem 594 (Ring localization universal property). \( \square \)

**Definition 602.** The field of rational algebraic functions \( D(\mathcal{X}) \) for the set of indeterminates \( \mathcal{X} \) over the integral domain \( D \) is the field of fractions of the polynomial ring \( D[\mathcal{X}] \).

Despite the name, elements of the field of fractions are not actually functions, but merely formal expressions. In particular, an analog of theorem 571 (Polynomial algebra universal property) does not really make sense.

**Proposition 603.** Let \( k \subseteq \mathbb{K} \) be fields and let \( A \) be an arbitrary subset of \( \mathbb{K} \).

Let \( k[A] \) be the ring obtained by adjoining the elements of \( A \) to \( k \) as described in proposition 576. The field of fractions of \( k[A] \) is the smallest field extension of \( k \) containing \( A \).

We denote this extension by \( k(A) \). It should not be confused with the image of the evaluation homomorphism on the field of rational functions; the rational functions do not actually have an evaluation homomorphism.

Proof. It follows from proposition 576 that \( k[A] \) is the smallest superring of \( k \) containing \( A \). By theorem 601 (Field of fractions universal property), \( k(A) \) is the smallest field containing \( k[A] \). \( \square \)
10.6. Modules

**Definition 604.** A **module** is a semimodule over a ring rather than a semiring.
This makes the identity law (192) redundant.

Modules have the following metamathematical properties:

(a) The first-order theory is identical to the theory of semimodules.

(b) A first-order homomorphism between two $R$-modules $M$ and $N$ is simply a linear map.

(c) The set $A \subseteq M$ is a submodule of $M$ if it is a sub-semimodule of $M$, i.e. a subgroup of $M$ that is closed under scalar multiplication. We say that $A$ is a submodule of $M$.

As a consequence of proposition 863, the image of a module homomorphism is a submodule of its range.

(d) The trivial module is the trivial pointed set $\{0\}$.

(e) A bimodule is simply a bisemimodule over a ring.

(f) For a fixed ring $R$, we denote the category of $\mathcal{U}$-small models by $\mathcal{U}\text{-Mod}_R$.

It is a very well-behaved category, even more than the category $\mathcal{U}\text{-Grp}$ of $\mathcal{U}$-small groups.

- The trivial module $\{0\}$ is a zero object. Therefore, we can define kernels and cokernels, and cokernels for modules are particularly simple.

- The free semimodules over a ring are modules, and theorem 552 (Free semimodule universal property) ensures that this is left adjoint to the forgetful functor $U : \mathcal{U}\text{-Mod}_R \rightarrow \mathcal{U}\text{-Set}$. Therefore, by proposition 917, the monomorphisms are exactly the injective homomorphisms, and that the categorical subobjects correspond to submodules.

- Every epimorphism in $\mathcal{U}\text{-Mod}_R$ is surjective. This will be proved in proposition 608. Along with corollary 466, this shows that the categorical quotient objects correspond to quotient modules, which we will define shortly.

(g) The kernel of an $R$-module homomorphism $\varphi : M \rightarrow N$ is its zero locus $\varphi^{-1}(0_N)$. This is a submodule of $M$. It is precisely the kernel of the underlying group in the sense of definition 455 (h), and the categorical kernel in the category of modules.

(h) The categorical cokernel of an $R$-homomorphism $\varphi : M \rightarrow N$ in the category $\text{Mod}_R$ is simply the additive quotient group $N/ \text{img} \varphi$. The quotient group is again a module over $R$ because $N$ is closed under scalar multiplication and, for every coset $x + N$,

$$r(x + N) = rx + rN = rx + N$$

is again a coset in $N/ \text{img} \varphi$.

In particular, given a submodule $N$ of $M$, we can form the quotient module $M/N$. In practice, quotients are conveniently characterized by theorem 606 (Quotient module universal property).
(i) Analogously to simple groups, if the only proper submodule of $R$ is the trivial module $\{0_M\}$, we say that $M$ is a **simple module**.

The trivial module itself is not simple, because it has no proper ideals.

**Proposition 605.** We have an isomorphism of categories $\text{Ab} \cong \text{Mod}_\mathbb{Z}$.

More concretely, every abelian group $G$ is a left module over $\mathbb{Z}$ with scalar multiplication given by recursively defined multiplication

$$
\cdot : \mathbb{Z} \times G \to G
$$

$$
n \cdot x := \begin{cases} 
0_G, & n = 0, \\
n \cdot x + x, & n > 1, \\
-(n \cdot x), & n < 1.
\end{cases}
$$

(212)

Conversely, in every module over $\mathbb{Z}$, scalar multiplication matches the recursively defined multiplication.

Compare this result to proposition 546.

**Proof.** Simple refinement of proposition 546.

**Theorem 606** (Quotient module universal property). For every $R$-module $M$ and every submodule $N$ of $M$, the quotient module $R/I$ has the following universal mapping property:

Every $R$-module homomorphism $\varphi : M \to K$ satisfying $N \subseteq \ker \varphi$ uniquely factors through $M/N$. That is, there exists a unique $R$-module homomorphism $\overline{\varphi} : M/N \to K$, such that the following diagram commutes:

$$
\begin{array}{ccc}
M & \xrightarrow{\varphi} & K \\
\downarrow{\pi} & & \downarrow{\overline{\varphi}} \\
M/N & \end{array}
$$

(213)

In the case where $N = \ker \varphi$, $\overline{\varphi}$ is an embedding.

Compare this result to theorem 464 (Quotient group universal property) and theorem 649 (Quotient algebra universal property).

**Proof.** Simple refinement of theorem 464 (Quotient group universal property).

**Theorem 607** (Quotient submodule lattice theorem). Given a submodule $N$ of $M$, the function $K \mapsto K/N$ is a lattice isomorphism between the lattice of submodules of $M$ containing $N$ and the lattice of submodules of the quotient $M/N$.

Compare this result to theorem 467 (Quotient subgroup lattice theorem) and theorem 651 (Quotient ideal lattice theorem).

**Proof.** Simple refinement of theorem 467 (Quotient subgroup lattice theorem).
**Proposition 608.** Every epimorphism in $\textbf{Mod}_R$ is surjective.

**Proof.** Let $\varphi : M \to N$ be an $R$-module epimorphism. Consider the canonical projection $\pi : N \to N/\text{img} \varphi$ and the zero morphism $z : N \to N/\text{img} \varphi$. Clearly

$$\pi \circ \varphi = z \circ \varphi,$$

and thus $\pi = z$ is the zero morphism.

By proposition 457 (h), $\ker \pi = \text{img} \varphi$, and since $\ker \pi = N$, it follows that $\varphi$ is a surjective function. \hfill $\square$

**Definition 609.** A **presentation** of the $R$-module $M$ is an epimorphism $\varphi : R^{\oplus A} \to M$, where $R^{\oplus A}$ is a **free semimodule**.

By theorem 606 (Quotient module universal property),

$$M = \text{img} \varphi \cong R^{\oplus A} / \ker \varphi.$$

Analogously to group presentations, we say that $M$ is finitely generated/related/presented if there exists an appropriate presentation.

**Definition 610.** Let $M$ be an $R$-module and fix a subset $B \subseteq M$. We say that the elements of $B$ are linearly independent if any of the following conditions hold:

(a) A linear combination in $B$ sums to zero if and only if it is trivial.

(b) The linear combination evaluation map

$$\Phi_B : R^{\oplus B} \to M$$

$$\Phi_B(\{t_b\}_{b \in B}) := \sum_{b \in B} t_b b$$

is injective.

Unsurprisingly, if the elements of $B$ are not linearly independent, we say that they are **linearly dependent**.

Compare this concept to algebraic dependence.

**Proof of correctness.**

**Proof that 610 (a) implies 610 (b).** Suppose that only the trivial linear combination of $B$ sums to zero. If $\sum_{b \in B} t_b b = \sum_{b \in B} r_b b$, then

$$\sum_{b \in B} t_b b - \sum_{b \in B} r_b b \overset{\text{(i)}}{=} \sum_{b \in B} (t_b - r_b) b = 0,$$

implying that $t_b = r_b$ for every $b \in B$.

Hence, $\Phi_B$ is injective.

**Proof that 610 (b) implies 610 (a).** Trivial. \hfill $\square$
Remark 611. Like all concepts related to linear combinations, linear dependence may behave differently depending on the underlying ring. For example, every irrational number is a linear combination of itself, but it is not a linear combination of rational numbers.

For simplicity, we will not specify the ring explicitly unless this may cause confusion.

**Proposition 612.** Linear (in)dependence in the $R$-module $M$ has the following basic properties:

(a) The zero vector $0_M$ is by itself linearly dependent.

(b) If $A$ is a linearly dependent set and $A \subseteq B$, then $B$ is also a linearly dependent set.

(c) If $B$ is a linearly independent set and $A \subseteq B$, then $A$ is also a linearly independent set.

(d) The set $A \cup \{x\}$ is linearly dependent if and only if $x \in \text{span } A$.

This may not hold for more general modules.

(e) For any set of vectors $A$, if $x \in \text{span } A \setminus A$, then $A \cup \{x\}$ is a linearly dependent set.

A partial converse is stated in proposition 625 (a).

**Proof.**

**Proof of 612 (a).** Trivial.

**Proof of 612 (b).** Trivial.

**Proof of 612 (c).** Trivial.

**Proof of 612 (e).** By proposition 554, there exists a linear combination of members of $A$ such that

$$x = \sum_{k=1}^{n} t_k x_k.$$ 

If $x = 0_M$, then $A \cup \{x\}$ is linearly dependent by proposition 612 (a) and proposition 612 (b). If $x \neq 0_M$, then $A \cup \{x\}$ is linearly dependent because $x$ is a nontrivial linear combination of other vectors of $A$. 

**Example 613.** We list several (counter)examples for linear dependence:

(a) The columns

$$
\begin{pmatrix}
0 \\
1 \\
0
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix}
\begin{pmatrix}
0 \\
1 \\
1
\end{pmatrix}
$$

**Definition 614.** Let $M$ be a left $R$-module and fix a subset $B \subseteq M$. We say that $B$ is a *Hamel basis* or simply *basis* of $M$ if any of the following equivalent conditions hold:

(a) It is a spanning set of linearly independent elements.
(b) The linear combination evaluation map

\[ \Phi_B : R^{\oplus B} \to M \]

\[ \Phi_B(\{ t_b \}_{b \in B}) := \sum_{b \in B} t_b b \]

is bijective.

(c) \( M \) is linearly isomorphic to the free module \( R^{\oplus B} \).

It is established terminology to say that \( M \) is a free module.

**Proof of correctness.**

**Proof that 614 (a) implies 614 (b).** Suppose that \( B \) is a spanning set of linearly independent elements.

Definition 610 (b) is satisfied, hence \( \Phi_B \) is injective.

Furthermore, \( B \) is spanning, meaning that \( \text{span} B = M \). By definition 610 (b), it is surjective.

**Proof that 614 (b) implies 614 (c).** If \( \Phi_B \) is bijective, then it is a linear isomorphism.

**Proof that 614 (c) implies 614 (a).** Suppose that \( \Psi : R^{\oplus B} \to M \) is a linear isomorphism. By definition 610 (b), \( \Psi \) is surjective, and hence a spanning set of \( M \). Furthermore, it satisfies definition 610 (b), meaning that \( B \) is a linearly independent set. \( \square \)

**Example 615.** As a \( \mathbb{Z} \)-module, the additive group of \( \mathbb{Q} \) has no basis.

Indeed, given any two rational numbers \( a/b \) and \( c/d \), we have

\[ cb \cdot \frac{a}{b} + da \cdot \frac{c}{d} = 0. \]

Therefore, only singleton sets of rational numbers are linearly independent with respect to \( \mathbb{Z} \). But no single integer generates \( \mathbb{Q} \).

**Proposition 616.** Bases of the \( R \)-module \( M \) have the following basic properties:

(a) Every linearly independent subset of \( M \) is a basis for its linear span.

(b) Every basis of \( M \) is maximal among linearly independent sets.

The converse holds for vector spaces — see proposition 625 (c).

(c) Every basis of \( M \) is minimal among spanning sets.

The converse holds for vector spaces — see proposition 625 (c).

**Proof.**

**Proof of 616 (a).** Trivial.

**Proof of 616 (b).** Follows from proposition 612 (e).
**Proof of 616 (c).** Let \( B \) be a basis of \( M \) and suppose that \( A \subseteq B \) is also a spanning set of \( M \). Then there exists some vector \( x \in B \setminus A \). Since both sets are spanning, \( \text{span} \ A = \text{span} \ B \). By proposition 612 (e), \( A \cup \{ x \} \) is linearly dependent, and by proposition 612 (c), \( B \) is linearly dependent. This contradicts the assumption that \( B \) has a basis.

Therefore, no proper subset of \( B \) is spanning for \( M \). \( \square \)

**Proposition 617.** Consider the \( \mathbb{R} \)-modules \( \{ M_k \}_{k \in \mathbb{K}} \). Suppose that \( \{ B_k \}_{k \in \mathbb{K}} \) are bases of the corresponding modules. Then the set

\[
B := \bigcup_{k \in \mathbb{K}} B_k
\]

is a basis for the direct sum

\[
M := \bigoplus_{k \in \mathbb{K}} M_k.
\]

**Proof.** Regard \( M_k \) as a subspace of the direct sum \( M \). The subspaces \( M_k \) and \( M_n \) are disjoint for \( k \neq n \). Thus, every vector from \( M \) is a unique sum of vectors from the subspaces.

Since, for every \( k \in \mathbb{K} \), every vector from \( M_k \) can further be uniquely represented as a linear combination of vectors from \( B_k \), we conclude that every vector from \( M \) is a linear combination of vectors from \( B \). \( \square \)

**Definition 618.** We say that the semimodule element \( x \) is a **torsion element** if there exists some nonzero scalar \( t \) such that \( t \cdot x \) is the zero vector. A semimodule without torsion elements is called **torsion-free**.

**Example 619.** We list examples of torsion elements:

(a) When regarded as a module over itself, every zero divisor of a ring is a torsion element. For example, 2 and 3 in \( \mathbb{Z}_6 \).

(b) The matrix

\[
\begin{pmatrix}
2 & 0 \\
0 & 3
\end{pmatrix}
\]

is a torsion element of the **matrix algebra** \( \mathbb{Z}_6^{2 \times 2} \) because \( 2 \cdot 3 = 3 \cdot 4 = 0 \).

**Proposition 620.** If the \( \mathbb{R} \)-module \( M \) has a basis, it is torsion-free.

**Proof.** Suppose that \( t \cdot x \) for some nonzero scalar \( t \) and nonzero vector \( x \). Let \( B \) be a basis of \( M \) and let \( x = \sum_{b \in B} \pi_b(x) \cdot b \) be the basis decomposition of \( x \). Then this is a nontrivial linear combination that sums to zero, which contradicts the assumption that \( B \) is a basis. \( \square \)

**Definition 621.** If \( B \) is a basis of the \( \mathbb{R} \)-module \( M \), the inverse of the linear isomorphism from definition 614 (b) is

\[
\pi_B : M \to \mathbb{R}^{\oplus B}
\]

\[
\pi_B \left( \sum_{b \in B} t_b \cdot b \right) := \{ t_b \}_{b \in B}.
\]
We denote the $b$-th component of this function by $\pi_b : M \to R$. This is also a linear map, which we call the \textbf{coordinate projection} for $b$.

For every vector $x$ in $M$, we thus have

$$x = \sum_{b \in B} \pi_b(x) \cdot b.$$ 

This linear combination is unique, and we call it the \textbf{decomposition} of $x$ along $B$.

As in \textbf{remark 553}, we sometimes take only the basis vectors with nonzero coefficients as “the” decomposition.

\textbf{Proof of correctness.} The decomposition is indeed unique — if

$$x = \sum_{b \in B} t_b b = \sum_{b \in B} r_b b,$$

then

$$\sum_{b \in B} (t_b - r_b) b = 0_M.$$ 

Since $B$ is a basis, it is linearly independent and hence $t_b = r_b$ for all $b$ in $B$. \hfill $\square$

\textbf{Definition 622.} Let $R$ be a \textbf{commutative ring} and let $\alpha$ be either a finite \textbf{ordinal} or the smallest infinite ordinal $\omega$. The \textbf{standard basis} of the \textbf{free module} $R^\alpha$ is the transfinite sequence $\{e_i\}_{j < \alpha}$ whose projection maps satisfy

$$\pi_i(e_j) = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases}$$

When $\alpha$ is finite, the vectors in $R^\alpha$ are finite sequences, and we often conflate them with \textbf{column vectors} or, less often, \textbf{row vectors}. See \textbf{remark 711} for a more detailed discussion of terminology.

\textbf{Proposition 623.} For a nontrivial commutative ring $R$, the set

$$\left\{ \prod_{X \in \mathcal{X}} X^{\gamma_X} \mid \gamma \text{ is a multi-index} \right\}$$

of all monomials is a \textbf{basis} for \textbf{polynomial ring} $R[\mathcal{X}]$.

\textbf{Proof.} In \textbf{definition 570}, we have defined a polynomial as a \textbf{free $R$-module} over the set of all monomials. \hfill $\square$

\textbf{Definition 624.} A \textbf{vector space} is a \textbf{left module} over a field.

We denote the category of vector spaces over $\mathbb{K}$ by $\textbf{Vect}_\mathbb{K}$.

\textbf{Proposition 625.} The \textbf{vector space} $V$ over $\mathbb{K}$ has the following basic properties:
(a) For a linearly independent set \( A \), if \( x \in V \setminus \text{span} A \), then \( A \cup \{x\} \) is also a linearly independent set.

A more general converse holds in proposition 612 (e).

(b) Every maximally linearly independent subset of \( V \) is a basis.

The converse holds more generally — see proposition 616 (b).

(c) Every minimal spanning subset of \( V \) is a basis.

The converse holds more generally — see proposition 616 (c).

(d) Given a subspace \( U \) of \( V \), if \( A \) is a finite basis of \( U \) and if \( V \) has a finite basis, then we can extend \( A \) to a finite basis of \( V \). That is, there exists a finite basis \( B \) of \( V \) such that \( A \subseteq B \).

(e) If \( \text{span}\{a_1, \ldots, a_n\} \subseteq \text{span}\{b_1, \ldots, b_m\} \), then \( n \leq m \).

Proof.

**Proof of 625 (a).** Suppose that \( x \in V \setminus \text{span} A \) and that \( A \cup \{x\} \) is linearly dependent, Then there exist coefficients \( t_0, t_1, \ldots, t_n \) such that

\[
t_0x + \sum_{k=1}^{n} t_k a_k = 0_M.
\]

If \( t_0 = 0_{K} \), then \( \sum_{k=1}^{n} t_k a_k \in \text{span} A \) and thus \( t_1 = \cdots = t_n = 0_{K} \).

Otherwise, we can divide by \( t_0 \) to obtain

\[
x = -\sum_{k=1}^{n} \frac{t_k}{t_0} a_k,
\]

which implies that \( x \in \text{span} A \), contradicting our choice of \( x \).

Therefore, \( A \cup \{x\} \) satisfies definition 610.

**Proof of 625 (b).** Let \( A \) be a maximally linearly independent set.

Suppose that it is not a spanning set and let \( x \in V \setminus \text{span} A \). By proposition 625 (a), the set \( A \cup \{x\} \) is linearly independent, contradicting the maximality of \( A \).

Therefore, \( A \) is a spanning set for \( V \).

**Proof of 625 (c).** Let \( A \) be a minimal spanning set. Suppose that it is linearly dependent. Then there exist distinct vectors \( x_1, \ldots, x_n \) in \( A \) and scalars \( t_1, \ldots, t_n \), at least one of which is nonzero, such that

\[
\sum_{k=1}^{n} t_k x_k = 0_M.
\]

Let \( k_0 \) be the smallest index such that \( t_{k_0} \neq 0_{R} \). Then

\[
x_{k_0} = -\sum_{k \neq k_0} \frac{t_k}{t_{k_0}} a_k.
\]

Then \( \text{span} A = \text{span} A \setminus \{x_{k_0}\} \), contradicting the minimality of \( A \).
Proof of 625 (d). Let $a_1, \ldots, a_n$ be a basis of $U$ and let $b_1, \ldots, b_m$ be a basis of $V$.

We use recursion on $k \leq n$ to build linearly independent sets of the form

$$\{a_1, \ldots, a_n, b_{i_1}, \ldots, b_{i_k}\},$$

at least one of which will be a basis.

The base case $k = 0$ is vacuous. Now suppose that the vectors $a_1, \ldots, a_n, b_{i_1}, \ldots, b_{i_k}$ are linearly independent. If there exists an index $i_{k+1}$ distinct from $i_1, \ldots, i_k$ such that $b_{i_{k+1}}$ is not in

$$L_k := \text{span}\{a_1, \ldots, a_n, b_{i_1}, \ldots, b_{i_k}\},$$

then, by proposition 625 (a), the vectors $a_1, \ldots, a_n, b_{i_1}, \ldots, b_{i_k}$ are linearly independent. Otherwise, every basis vector $b_1, \ldots, b_m$ belongs to $L_k$, hence it is a spanning set of $V$ of linearly independent vectors.

Since the basis $b_1, \ldots, b_n$ of $V$ necessarily belongs to $L_n$, it follows that $L_k$ will be a basis of $V$ for some $k \leq n$.

Proof of 625 (e). We will use induction on $n$. The base case $n = 0$ holds because $\{0_V\}$ is the only submodule of the linear span of zero vectors.

Suppose that the statement holds for $n - 1$ and note that $b_m$ can be decomposed as

$$b_m = \sum_{k=1}^{n} t_k a_k.$$ 

Then

$$a_k = \frac{1}{t_{k_0}} b_m - \sum_{k \neq k_0} \frac{t_k}{t_{k_0}} a_k.$$ 

Thus,

$$\text{span}\{a_1, \ldots, a_n\} = \text{span}\{a_1, \ldots, a_{k_0-1}, a_{k_0+1}, \ldots, a_n, b_m\}.$$ 

We can thus remove $b_m$ to obtain the inclusion

$$\text{span}\{b_1, \ldots, b_{m-1}\} \subseteq \text{span}\{a_1, \ldots, a_{k_0-1}, a_{k_0+1}, a_n\}.$$ 

From the inductive hypothesis, we conclude that $m - 1 \leq n - 1$, and hence $m \leq n$.

Theorem 626 (Vector space basis existence). Every vector space has a basis.

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (m).

Proof.

Proof that Zorn’s lemma implies vector space existence. Let $V$ be a vector space over $\mathbb{K}$. Let $\mathcal{B}$ be the family of all linearly independent subsets of $V$.

The family $\mathcal{B}$ is nonempty since any singleton from $V$ belongs to $\mathcal{B}$. The union of any chain $\mathcal{B}' \subseteq \mathcal{B}$ can then contain only linearly independent elements since otherwise we would have that some set in $\mathcal{B}'$ is not linearly independent. Thus, theorem 1240 (Zorn’s lemma) shows the existence of a maximal linearly independent set $\mathcal{B}$. By proposition 625 (b), $\mathcal{B}$ is a basis.
Proof that vector space existence implies the axiom of choice. Shown in [Bla84]. □

Proposition 627. All bases of a vector space are equinumerous.

We define the dimension \( \dim V \) of the vector space \( V \) as the cardinality of any of its bases. By theorem 626 (Vector space basis existence), every vector space has a dimension.

Proof. Let \( A \) and \( B \) be bases of \( V \).

Proof for finite bases. Suppose that \( a_1, \ldots, a_n \) are the vectors of \( A \) and, aiming at a contradiction, suppose that \( B \) is infinite.

Every vector \( a_i \) of \( A \) can be decomposed along \( B \) as

\[
a_i = \sum_{j=1}^{m_i} t_{ij}^{(j)} b_{j}^{(i)}.
\]

for appropriate scalars from \( \mathbb{K} \) and vectors from \( B \). Since this can be done for every \( a_i \) in \( A \), we conclude that

\[
V = \text{span}(a_1, \ldots, a_n) = \text{span}(b_1^{(1)}, \ldots, b_{m_1}^{(1)}, \ldots, b_1^{(n)}, \ldots, b_{m_n}^{(n)}).
\]

Hence, a finite subset of \( B \) spans \( V \), contradicting proposition 616 (c). Therefore, \( B \) must be a finite set.

Let \( b_1, \ldots, b_m \) be the vectors of \( B \). By applying proposition 625 (e) twice, we conclude that \( n \leq m \) and \( m \leq n \), hence \( n = m \).

Proof for infinite bases. Suppose that both \( A \) and \( B \) are infinite. Let \( S_x \) be the set of vectors in \( B \) with nonzero coefficients in the decomposition of \( x \in V \).

- \( S_x \) is necessarily a finite set.
- Every vector \( x \) in \( V \) belongs to \( S_a \) for some \( a \in A \).
  
  Indeed, \( x \) can be decomposed along \( A \) as

\[
x = \sum_{a \in A} \pi_a(x) \cdot a,
\]

and \( x \in S_{\pi(a)} \) whenever \( \pi(a) \neq 0_{\mathbb{K}} \).
- For every basis vector \( b \) in \( B \), \( S_b = \{ b \} \).
- We have

\[
B \subseteq \bigcup_{a \in A} S_a
\]

Therefore,

\[
\text{card}(B) \leq \text{card} \left( \bigcup_{a \in A} S_a \right) \leq \text{card} \left( \prod_{a \in A} S_a \right) \leq \text{card}(A \times \omega) \leq \text{card}(A) \cdot \aleph_0 \overset{\text{1043 (b)}}{\leq} \text{card}(A).
\]

Since \( A \) and \( B \) were arbitrary bases, we can exchange them to obtain the converse inequality, and thus \( \text{card}(A) = \text{card}(B) \). □
**Proposition 628.** All bases of a module over a nontrivial commutative ring are linearly isomorphic.

We define the **rank** of a module as the cardinality of any of its bases. This is a generalization of vector space dimensions. Unlike vector space dimensions, however, module ranks may not exist — see example 615. Modules with at least one basis are often called **free**, however we will prefer to use the term for modules with a concrete bases, as in definition 551.

**Proof.** Suppose that $A$ and $B$ are bases of the $R$-module $M$. Let $I$ be a maximal ideal of $R$.

By corollary 652, $𝕂 = R/I$ is a field (this is a forward reference to corollary 652). Given an isomorphism $Φ : R^⊕A → R^⊕B$ with components $Φ_b : R^⊕A → R$, define

$$Ψ : 𝕂^⊕A → 𝕂^⊕B$$

$$Ψ([t_a + I]_{a ∈ A}) := \left\{Φ_b([t_a]_{a ∈ A}) + I\right\}_{b ∈ B}$$

This is clearly an isomorphism if it is well-defined. Indeed, suppose we are given $t_a + I = t'_a + I$ for every $a ∈ A$, i.e. $t_a - t'_a ∈ I$. Then

$$Φ_b([t_a]_{a ∈ A}) - Φ_b([t'_a]_{a ∈ A}) = Φ_b([t_a - t'_a]_{a ∈ A}) = Φ_b([t_a - t'_a]_{a ∈ A}) = \sum_{a ∈ A} \left(t_a - t'_a\right)Φ_b(a) ∈ I.$$ 

Thus,

$$Φ_b([t_a]_{a ∈ A}) + I = Φ_b([t'_a]_{a ∈ A}) + I.$$ 

Therefore, $Ψ$ is a well-defined linear isomorphism. Applying proposition 627, we conclude that $A$ and $B$ are equinumerous.

**Corollary 629.** For a nontrivial commutative ring $R$, if a finitely-generated $R$-module has a basis, the basis is finite.

**Proof.** Follows from proposition 628.

**Example 630.** For a nontrivial commutative ring $R$, every ideal of $R$ is a submodule of the rank-one module $R$. For example, both $ℤ$ and $2ℤ$ are $ℤ$-modules of rank one.

For a field $𝕂$, every ideal of $𝕂$ is a submodule of the unidimensional vector space $𝕂$. The only ideals of $𝕂$ are the zero ideal, whose dimension is zero, and the field itself, whose dimension is one.

**Lemma 631.** For vector spaces $U ⊆ V$ we have rank $U ≤$ rank $V$.

**Proof.** Let $A$ be a basis of $U$. By proposition 625(d), there exists a basis $B$ of $V$ such that $A ⊆ B$. Then clearly rank $U ≤$ rank $V$.

**Proposition 632.** Let $U$ be a subspace of the finite-dimensional vector space $V$. Then

$$V ≅ U ⊕ (V/U).$$
Proof. By lemma 631, \( \dim U \leq \dim V \). Let \( a_1, \ldots, a_m \) be an ordered basis for \( U \). By proposition 625 (d), there exist vectors \( a_{m+1}, \ldots, a_n \) such that \( a_1, \ldots, a_n \) is a basis of \( V \). We will show that \( a_{m+1} + U, \ldots, a_n + U \) is a basis of \( V/U \), thus demonstrating the isomorphism.

First, we must show that it is a spanning set. Fix a vector \( x + U \) from \( V/U \). We know that \( x \) can be decomposed as

\[
x = \sum_{k=1}^{n} \pi_k(x) \cdot a_k.
\]

Since \( a_1, \ldots, a_m \) belong to \( U \),

\[
x + U = \sum_{k=m+1}^{n} \pi_k(x) \cdot a_k + U.
\]

Hence, \( x + U \) can be represented as a linear combination of the vectors \( a_{m+1} + U, \ldots, a_n + U \).

To show uniqueness, suppose that

\[
x + U = \sum_{k=m+1}^{n} t_k(a_k + U) = \sum_{k=m+1}^{n} r_k(a_k + U).
\]

Then

\[
U = \sum_{k=m+1}^{n} (t_k - r_k)(a_k + U).
\]

Since neither of \( a_{m+1}, \ldots, a_n \) are in \( U \), it follows that \( t_k = r_k \) for \( k = m + 1, \ldots, n \).

Therefore, \( a_{m+1} + U, \ldots, a_n + U \) is a basis of \( V/U \).

Proposition 633. Assuming that the \( R \)-modules \( M_1, \ldots, M_n \) have bases, the rank of their direct sum

\[
M_1 \oplus \cdots \oplus M_n
\]

is the sum of cardinals

\[
\text{rank} M_1 + \cdots + \text{rank} M_n.
\]

Proof. Follows from proposition 617.

Theorem 634 (Rank-nullity theorem). For every linear map \( \varphi : U \to V \) between finite-dimensional vector spaces, we have

\[
U \cong \ker \varphi \oplus \text{img} \varphi.
\]

In particular,

\[
\dim U = \dim \ker \varphi + \dim \text{img} \varphi.
\]  

The dimension of the kernel of \( \varphi \) is often called the nullity of \( \varphi \) and the dimension of the image - the rank of \( \varphi \).
Proof. By theorem 606 (Quotient module universal property),

\[
\text{img } \varphi \cong U / \ker \varphi.
\]

By proposition 632,

\[
U \cong \ker \varphi \oplus (U / \ker \varphi) \cong \ker \varphi \oplus \text{img } \varphi.
\]

The equality (634) then follows from proposition 633. \qed
10.7. Univariate polynomials

We will discuss here the polynomial ring $R[X]$ in one indeterminate over the nontrivial commutative unital ring $R$. We call them univariate polynomials using the general convention for function arguments from definition 966 (c). Remark 572 discusses why we often focus only on finitely many indeterminates, and why the theory of univariate polynomials is often sufficient.

Polynomials are not functions in general, and the exact relationship between polynomials and polynomial functions is discussed in theorem 571 (Polynomial algebra universal property) and proposition 699.

Definition 635. The degree of the nonzero univariate monomial $X^k$ is its power $k$. More generally, the degree of a multivariate monomial $\prod_{X \in \mathcal{X}} X^{\gamma X}$ in the set of indeterminates $\mathcal{X}$ is the multi-index norm $\|\gamma\| = \sum_{X \in \mathcal{X}} \gamma X$.

The degree $\deg(p)$ of a polynomial $p$ is the maximal degree of its nonzero monomials. For the zero polynomial, we leave the degree undefined.

For a univariate polynomial

$$p(X) = \sum_{k=0}^{\infty} a_k X^k = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots,$$

the degree $n := \deg(p)$ allows us to write

$$p(X) = \sum_{k=0}^{n} a_k X^k = a_0 + a_1 X + a_2 X^2 + \cdots + a_{n-1} X^{n-1} + a_n X^n.$$

This notation also subsumes the zero polynomial. We call $a_n$ the leading coefficient and $a_0$ the constant coefficient of the polynomial.

We introduce the following names for univariate polynomials of certain degrees:

<table>
<thead>
<tr>
<th>Degree Type</th>
<th>$\deg(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$0$ or $p$ is the zero polynomial</td>
</tr>
<tr>
<td>Linear</td>
<td>$1$</td>
</tr>
<tr>
<td>Quadratic</td>
<td>$2$</td>
</tr>
<tr>
<td>Cubic</td>
<td>$3$</td>
</tr>
<tr>
<td>Quartic</td>
<td>$4$</td>
</tr>
<tr>
<td>Quintic</td>
<td>$5$</td>
</tr>
</tbody>
</table>

Definition 636. We say that the nonzero univariate polynomial $p(X)$ is monic if its leading coefficient is 1.

Proposition 637. The polynomial degree has the following basic properties:

(a) For any two zero polynomials satisfying $p(X) \neq -q(X)$, we have

$$\deg(p + q) \leq \max\{\deg p, \deg q\}.$$ (216)
For any two nonzero polynomials \( p(X) \) and \( q(X) \) whose leading coefficients do not multiply to zero, we have
\[
\deg(pq) = \deg p + \deg q.
\] (217)

An easy sufficient condition for (217) is for the ring to be entire, although it is also sufficient for the ring to be nontrivial (so that \( 0_R \neq 1_R \)) and either \( p(X) \) or \( q(X) \) to be monic.

**Proof.** Fix nonzero polynomials
\[
p(X) := \sum_{k=0}^{n} a_k X^k, \\
q(X) := \sum_{k=0}^{m} b_k X^k.
\]

**Proof of 637 (a).** Additionally assume that \( p(X) \neq -q(X) \) since otherwise \( p(X) + q(X) = 0 \) and \( \deg(p + q) \) is undefined. Thus, there exists at least one index \( k = 1, 2, \ldots \), so that \( a_k \neq b_k \). Denote by \( k_0 \) the largest such index (only finitely many are nonzero). Then
\[
a_k = b_k = 0 \text{ for } k > k_0.
\]

Therefore, \( \deg(p + q) = k_0 \). Note that \( k_0 \) cannot exceed both \( \deg p \) and \( \deg q \) because it corresponds to a nonzero coefficient. Thus, \( k_0 \leq \max\{\deg p, \deg q\} \).

**Proof of 637 (b).** The coefficient \( c_{n+m} \) of the product \( p(X)q(X) \) is \( a_nb_m \) by definition. By assumption, it is nonzero. Then, since \( c_{n+m+1} = 0 \), we have
\[
\deg(pq) = \deg p + \deg q.
\]

\[\square\]

**Algorithm 638** (Euclidean division of polynomials). Fix two univariate polynomials \( f(X) \) and \( g(X) \), and assume that \( g(X) \) is monic.

We will build polynomials \( q(X) \) and \( r(X) \), where \( r(X) \) is either zero or \( \deg r < \deg g \), such that
\[
f(X) = g(X)q(X) + r(X).
\]
The algorithm only demonstrates existence; we will prove uniqueness right after it.

(a) If \( \deg f = \deg g = 0 \), necessarily \( g(X) = 1_R \), and in this case we define
\[
q(X) := f(X), \\
r(X) := 0_R.
\]

(b) If \( f(X) \) is the zero polynomial or \( \deg f < \deg g \), define
\[
q(X) := 0_R, \\
r(X) := f(X).
\]

In this case, \( r(X) \) is either zero or \( \deg r = \deg f < \deg b \).
Suppose that
\[ f(X) = a_nX^n + \hat{f}(X), \]
\[ g(X) = X^m + \hat{g}(X), \]
where \( n \) and \( m \) are positive, \( \hat{f}(X) \) is either zero or \( \deg \hat{f} < \deg f \), and similarly for \( \hat{g}(X) \).

Then
\[ f(X) - g(X)a_nX^{n-m} = a_nX^n + \hat{f}(X) - (b_mX^m + \hat{g}(X))a_nX^{n-m} = \]
\[ = a_nX^n + \hat{f}(X) - a_nX^n - \hat{g}(X)a_nX^{n-m} = \]
\[ = \frac{f(X) - \hat{g}(X)a_nX^{n-m}}{\hat{r}(X)}. \]

The polynomial \( \hat{r}(X) \) is either zero, in which case we define \( r(X) := \hat{r}(X) \), or \( \deg \hat{r} \leq n - 1 \).

In the latter case, we use the algorithm recursively to divide \( \hat{r}(X) \) by \( g(X) \), and obtain \( \hat{q}(X) \) and \( r(X) \) such that
\[ \hat{r}(X) := g(X)\hat{q}(X) + r(X), \]
where \( r(X) \) is either zero or \( \deg r < \deg g \).

Then
\[ \hat{r}(X) = f(X) - g(X)a_nX^{n-m} \]
\[ g(X)\hat{q}(X) + r(X) = f(X) - g(X)a_nX^{n-m} \]
\[ g(X)(\hat{q}(X) - a_nX^{n-m}) + r(X) = f(X). \]

Define
\[ q(X) := \hat{q}(X) - a_nX^{n-m}. \]

We have obtained polynomials \( r(X) \) and \( q(X) \) where \( r(X) \) is either zero or \( \deg r < \deg g \).

**Proof of correctness.**

**Proof of uniqueness.** Suppose that
\[ a(X) = g(X)q(X) + r(X) = g(X)\hat{q}(X) + \hat{r}(X), \]
where \( r(X) \) and \( \hat{r}(X) \) are either zero or have degree less than \( g(X) \).
Assume that \( r(X) \neq \hat{r}(X) \).

- If both \( r(X) \) and \( \hat{r}(X) \) are nonzero, we have
\[ g(X)(q(X) - \hat{q}(X)) = -(r(X) - \hat{r}(X)). \]
Since \( g(X) \) is monic and its leading coefficient \( 1_R \) is not a zero divisor, proposition 637 (b) holds, and thus
\[
\deg g + \deg(q - \tilde{q}) \overset{(217)}{=} \deg(g(q - \tilde{q})) = \deg(r - \tilde{r}) \overset{(216)}{=} \max\{\deg r, \deg \tilde{r}\} < \deg g,
\]
which is a contradiction.

- If \( r(X) \) is zero but \( \tilde{r}(X) \) is not, then
\[
g(X)q(X) = g(X)\tilde{q}(X) + \tilde{r}(X),
\]
implying that
\[
\tilde{r}(X) = g(X)(q(X) - \tilde{q}(X)).
\]
By (637 (b)), \( \deg g \leq \tilde{r} \), which contradicts our choice of \( \tilde{r}(X) \).

\[\square\]

**Definition 639.** Generalizing definition 204 from analysis, we define the **algebraic derivative** of a univariate polynomial
\[
p(X) = \sum_{k=0}^{n} a_k X^k = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_2 X^2 + a_1 X + a_0
\]
as
\[
p'(X) := \sum_{k=1}^{n} k a_k X^{k-1} = n a_n X^{n-1} + (n - 1)a_{n-1}X^{n-2} + \cdots + a_2 X + a_1.
\]

Via **natural number recursion**, we can define algebraic derivatives of order \( m \) as
\[
p^{(m)}(X) := \begin{cases} p(X) & m = 0 \\ (p^{(m-1)})'(X) & m > 0 \end{cases}
\]

**Proposition 640.** Algebraic derivatives have the following basic properties:

(a) The derivative operator \( p(X) \mapsto p'(X) \) is linear.

(b) \( p^{(n)}(X) \) is either zero or has degree \( \deg p - n \).

(c) We have the product rule
\[
(pq)' = p'q + pq'.
\]

(d) Leibniz’ rule holds:
\[
(pq)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} p^{(k)} q^{(n-k)}
\]

(e) If \( m \leq n \), the \( m \)-th derivative of \((X - u)^n\) is \( \frac{n!}{(n-m)!}(X - u)^{n-m} \).
Proof.

**Proof of 640 (a).** Trivial.

**Proof of 640 (b).** Trivial.

**Proof of 640 (c).** By proposition 640 (a), it is enough to consider the case where both \( p(X) \) and \( q(X) \) are monomials.

\[
p'(X)q(X) + p(X)q'(X) = n a_n X^{n-1} \cdot b_m X^m + a_n X^n \cdot m b_m X^{m-1} =
\]

\[
= (n + m) a_n b_m X^{n+m-1} =
\]

\[
= (a_n b_m X^{n+m})' =
\]

\[
= (pq)'(X).
\]

**Proof of 640 (d).** The proof in theorem 59 relies only on the product rule, hence it holds here as well.

**Proof of 640 (e).** We use outer induction on \( m \) and inner induction on \( n \).

The case \( m = n = 1 \) is obvious. Assume that the statement holds for \( m = 1 \) and \( n - 1 \). Then

\[
\left((X-u)^n\right)' = \left((X-u)^{n-1} \cdot (X-u)\right)' \overset{(19)}{=} \left((X-u)^{n-1}\right) (X-u) + (X-u)^{n-1} \overset{\text{ind.}}{=} n(X-u)^{n-1}.
\]

Now suppose that the statement holds for derivatives of order less than \( m \) and for every \( n \geq m \). Then,

\[
\left((X-u)^n\right)^{(m)} = \left(\left((X-u)^n\right)^{(m-1)}\right)^\circ \overset{\text{ind.}}{=} \left(\frac{n!}{(n-m+1)!} (X-u)^{n-m+1}\right)^\circ \overset{\text{ind.}}{=} \frac{n!}{(n-m)!} (X-u)^{n-m}.
\]

\( \square \)

**Definition 641.** Let \( R \) be a nontrivial commutative ring.

We say that the value \( u \in R \) is a root of multiplicity \( m \) for the univariate polynomial \( p(X) \in R[X] \) of degree \( n \geq m \) if any of the following equivalent conditions hold:

(a) The polynomial \( (X-u)^m \) divides \( p(X) \).

(b) The value \( u \) is a zero of the algebraic derivatives \( p^{(0)}(X), p^{(1)}(X), \ldots, p^{(m-1)}(X) \) of \( p(X) \).

Every polynomial \( p(X) \) has a multiset of roots.

**Proof of correctness.**

**Proof that 641 (a) implies 641 (b).** Suppose that \( (X-u)^m \) divides \( p(X) \). Then there exists a polynomial \( q(X) \) such that

\[
p(X) = (X-u)^m q(X).
\]
For the $n < m$-th derivative of $p(X)$, by proposition 640 (d), we have
\[
p^{(n)}(X) = \sum_{k=0}^{n} \binom{n}{k} (X-u)^{m} \binom{m}{k} q^{(n-k)}(X).
\]

Let $\Phi_u : R[X] \to R$ be the evaluation homomorphism at $u$. Then
\[
\Phi_u(p^{(n)}) = \sum_{k=0}^{n} \binom{n}{k} \frac{m!}{k!} 0_{R}^{m-k} \Phi_u(q^{(n-k)}(X)).
\]

For $n < m$, clearly $\Phi_u(p^{(n)}) = 0_R$.

**Proof that 641 (b) implies 641 (a).** Suppose that $u$ is a root of $p^{(0)}(X), \ldots, p^{(m-1)}(X)$. We will use induction on $m$ to show that $(X-u)^m \mid p(X)$.

The case $m = 0$ is trivial. Suppose that $u$ being a root of $p^{(0)}(X), \ldots, p^{(m-1)}(X)$ implies that $(X-u)^{m-1} \mid p(X)$, and additionally let $u$ be a root of $p^{(m)}(X)$.

By the inductive hypothesis, there exists a polynomial $q(X)$ such that
\[
p(X) = (X-u)^{m-1}q(X).
\]

By proposition 640 (d),
\[
p^{(m-1)}(X) = \sum_{k=0}^{m-1} \binom{m-1}{k} (X-u)^{m-1} \binom{m-1}{k} q^{(m-k-1)}(X).
\]

Then
\[
\Phi_u(p^{(m-1)}) = \sum_{k=0}^{m-1} \binom{m-1}{k} \frac{(m-1)!}{k!} 0_{R}^{m-k-1} \Phi_u(q^{(m-k-1)}(X)).
\]

All terms on the right are zero except for $\Phi_u(q)$. But $u$ is a root of $p^{(m-1)}$, implying that it is also a root of $q$.

We use algorithm 638 (Euclidean division of polynomials) to obtain a polynomial $s(X)$ and a constant polynomial $r(X) = r_0$ so that
\[
q(X) = (X-u)s(X) + r_0.
\]

Since $\Phi_u(q) = 0_R$, then necessarily $r_0 = 0_R$. Therefore,
\[
p(X) = (X-u)^{m-1}q(X) = (X-u)^m s(X).
\]

**Proposition 642.** Given monic polynomial $g(X)$ in a nontrivial commutative ring $R$, every coset in $R[X]/\langle g(X) \rangle$ has a unique representative that is either the zero polynomial or a polynomial of degree less than $g(X)$.
Proof. Let \( f(X) \) be an arbitrary polynomial. **Algorithm 638** (Euclidean division of polynomials) gives us polynomials \( q(X) \) and \( r(X) \) so that
\[
f(X) = g(X)q(X) + r(X),
\]
where \( r(X) \) is either zero or has degree less than \( g(X) \).

Multiples of \( q(X) \) are congruent to \( 0_R \) modulo the ideal \( (q(X)) \), hence \( f(X) \) is congruent to \( r(X) \).

By the uniqueness of \( r(X) \), the statement of the corollary follows. \( \square \)

**Remark 643.** Fix arbitrary commutative rings \( R \subseteq S \) and some element \( u \) of \( S \). By **theorem 571** (Polynomial algebra universal property), there exists a unique \( R \)-algebra homomorphism \( \Phi_u : R[X] \rightarrow S \) sending \( X \) to \( u \). If the kernel of \( \Phi_u \) is a principal ideal, and if \( p(X) \) is a generator, then we have the algebra isomorphism
\[
R[X]/(p(X)) \cong R[u].
\]

By **proposition 642**, there exists a correspondence between polynomials
\[
f(X) = \sum_{k=0}^{n} a_k X^k
\]
of degree less than \( \deg p \), and elements of the form
\[
\sum_{k=0}^{n} a_k u^k
\]
Furthermore, multiplication in \( R[u] \) corresponds to polynomial multiplication modulo \( p(X) \).

**Example 644.** The **Gaussian integers** are complex numbers \( z = a + bi \) with integer real and imaginary components. We can define several isomorphic rings for the Gaussian integers, demonstrating remark 643.

(a) We can take the quotient ring \( \mathbb{Z}[X]/(X^2 + 1) \). By **proposition 642**, the remainder from **algorithm 638** (Euclidean division of polynomials) can be used as a canonical representative within the quotient. The remainder must be either a constant or a linear polynomial. That is, \( r(X) = aX + Y \).

In order to make sense of the imposed ring structure in the quotient, we can see how multiplication modulo \( X^2 + 1 \) works. We have
\[
(bX + a)(dX + c) \equiv bdX^2 + (ad + bc)X + ac \pmod{X^2 + 1}
\]
\[
\equiv bd(X^2 + 1) + ((ad + bc)X - bd + ac) \pmod{X^2 + 1}
\]
\[
\equiv (ad + bc)X + (ac - bd) \pmod{X^2 + 1}.
\]
This is precisely the definition of multiplication of complex numbers as given in definition 40. Thus, 
\[ \mathbb{Z}[X]/(X^2 + 1) \]
is the desired ring of Gaussian integers.

(b) We can also adjoin \( i \) to \( \mathbb{Z} \) to obtain the ring \( \mathbb{Z}[i] \).

Given a Gaussian integer \( z = a + bi \), it corresponds to the polynomial 
\[ p_z(X) := a + bX. \]

Conversely, consider the evaluation homomorphism \( \Phi_i : \mathbb{Z}[X] \to \mathbb{C} \) for the imaginary unit. Let \( p(X) \in \mathbb{Z}[X] \). Then 
\[ p(i) = \Phi_i(p) = \sum_{k=0}^n a_k i^k = \sum_{\text{rem}(k,4)=0}^n a_k - \sum_{\text{rem}(k,4)=2}^n a_k + i \left( \sum_{\text{rem}(k,4)=1}^n a_k - \sum_{\text{rem}(k,4)=3}^n a_k \right). \]
This is clearly a Gaussian integer.

It remains to show that multiplication in \( \mathbb{Z}[i] \) is compatible with multiplication in \( \mathbb{C} \).

But complex multiplication is defined to be compatible with the notation \( a + bi \), that is, 
\[ (a + bi)(c + di) = ac + i bc + i ad - bd = (ac - bd) + i(bc + ad). \]

Thus, the Gaussian integers are precisely the homomorphic image of \( \mathbb{Z}[X] \) under \( \Phi_i \).

Corollary 645. For two nonzero monic polynomials \( p(X) \) and \( q(X) \) of the same degree, the quotient rings \( R[X]/\langle p(X) \rangle \) and \( R[X]/\langle q(X) \rangle \) are isomorphic as \( R \)-modules, but may not be isomorphic as \( R \)-algebras.

Proof. By proposition 642, for every coset in the quotient, algorithm 638 (Euclidean division of polynomials) gives us a unique representative of the corresponding degree. Addition and scalar multiplication must be the same in both.

As shown in example 644 and example 646, however, the vector multiplication operation may differ. \( \square \)

Example 646. Similarly to how the Gaussian integers were defined in multiple ways in example 644, remark 643 gives us an isomorphism
\[ \mathbb{Z}[X]/(X^2 - 2) \cong \mathbb{Z}[\sqrt{2}]. \]

The gist of this example is that, even though \( \mathbb{Z}[\sqrt{2}] \) and \( \mathbb{Z}[i] \) are isomorphic as modules, their vector multiplication operation is different. Indeed, multiplication modulo \( X^2 - 2 \) works as follows:
\[
(aX + b)(cX + bd) \cong acX^2 + (bc + ad)X + bd \\
\cong ac(X^2 - 2) + (bc + ad)X + 2ac + bd \quad \text{(mod } X^2 - 2 \text{)} \cong \\
\cong (bc + ad)X + (2ac + bd) \quad \text{(mod } X^2 - 2 \text{)}. 
\]
10.8. Algebras over rings

Definition 647. An algebra over a commutative ring $R$ rather than over a semiring exhibits some more interesting metamathematical properties.

(a) The first-order theory is identical to the theory of algebras over semimodules.

(b) A first-order homomorphism between two $R$-algebras $A$ and $B$ is a linear map that preserves multiplication. This is the same as for semirings.

(c) The set $A \subseteq M$ is a submodel of $M$ if it is a submodule of $M$ that is closed under algebra multiplication.

As a consequence of proposition 863, the image of an $R$-algebra homomorphism is a subalgebra of its range.

(d) The trivial $R$-algebra is the trivial pointed set $\{0_R\}$.

(e) For a fixed ring $R$, we denote the category of $\mathcal{U}$-small models by $\mathcal{U}\text{-Alg}_R$. It is concrete with respect to both $\mathcal{U}\text{-CRing}$ and $\mathcal{U}\text{-Mod}_R$.

Unfortunately, these categories are not as well-behaved as categories of modules.

(f) The kernel of an $R$-algebra homomorphism $\varphi : M \rightarrow N$ is its zero locus $\varphi^{-1}(0_N)$.

The kernel of a homomorphism is a both a submodule and a subring of $M$. It is the kernel of the underlying group, ring and module, and the categorical kernel in the category of modules.

(g) The categorical cokernel of an $R$-homomorphism $\varphi : M \rightarrow N$ in the category $\text{Alg}_R$ is simply quotient ring $N / \text{img} \varphi$.

More generally, let $I$ be a two-sided ideal of $M$. We identify each member $x$ of $R$ with its embedding $x \cdot 1_R$. Since $I$ is closed under multiplication with elements of $M$, it is also closed under multiplication with elements of $R$. Thus, $I$ is a submodule of the $R$-module $M$.

Thus, given an ideal $I$ of $M$, $M / I$ is both a quotient ring and a quotient module.

Proposition 648. The categories Ring of rings and $\text{Alg}_\mathbb{Z}$ of integer algebras are isomorphic. Compare this result to proposition 605.

Proof. Follows from proposition 546 by noting that, as in the proof of proposition 566, distributivity implies bilinearity. □

Theorem 649 (Quotient algebra universal property). For every unital $R$-algebra $M$ and ideal $I$ of $M$, the quotient algebra $M / I$ has the following universal mapping property:
Every ring homomorphism \( \varphi : M \rightarrow N \) satisfying \( I \subseteq \ker \varphi \) uniquely factors through \( M/I \). That is, there exists a unique homomorphism \( \tilde{\varphi} : M/I \rightarrow N \), such that the following diagram commutes:

\[
\begin{array}{c}
M \\
\xrightarrow{\pi} \\
M/I \\
\end{array} \hspace{0.5cm} \xrightarrow{\varphi} \hspace{0.5cm} \begin{array}{c} N \\
\end{array} \hspace{0.5cm} \begin{array}{c} \xrightarrow{\tilde{\varphi}} \\
\end{array} \\
\]

(220)

In the case where \( I = \ker \varphi \), \( \tilde{\varphi} \) is an embedding.

Compare this result to theorem 464 (Quotient group universal property) and theorem 606 (Quotient module universal property).

**Proof.** Simple refinement of theorem 464 (Quotient group universal property).

**Definition 650.** For a commutative ring \( R \), a **presentation** of the \( R \)-algebra \( M \) is a surjective homomorphism \( \varphi : R[S] \rightarrow M \) (epimorphisms may be too general), where \( R[S] \) is a polynomial ring with indeterminates \( S \).

By theorem 649 (Quotient algebra universal property), \( M = \text{img} \varphi \cong R[S]/ \ker \varphi \).

Analogously to group presentations, we say that \( M \) is finitely generated/related/presented if there exists an appropriate presentation.

**Theorem 651** (Quotient ideal lattice theorem). Given a two-sided ideal \( I \) of the unital \( R \)-algebra \( M \), the function \( N \mapsto N/I \) is a lattice isomorphism between the lattice of subalgebras of \( M \) containing \( I \) and the lattice of subalgebras of the quotient \( M/I \).

Furthermore, the subsets of prime, maximal or radical ideals are order-isomorphic.

Note that, by proposition 648, usual rings can be regarded as \( Z \)-algebras.

Compare this result to theorem 467 (Quotient subgroup lattice theorem) and theorem 607 (Quotient submodule lattice theorem).

**Proof.**

**Proof for general ideals.** Simple refinement of theorem 467 (Quotient subgroup lattice theorem).

**Proof for prime ideals.** Suppose that \( P \) is a prime ideal in \( M \) containing \( I \). Let \( A \) and \( B \) be ideals of \( P \) containing \( I \), such that \( (A/I)(B/I) \subseteq P/I \). By the general ideal lattice isomorphism, \( AB \subseteq P \), which implies that \( A \subseteq P \) or \( B \subseteq P \). Again using the general isomorphism, we conclude that \( A/I \subseteq P/I \) or \( B/I \subseteq P/I \), meaning that \( P/I \) is a prime ideal in \( M/I \).
Proof of necessity. Suppose that $P/I$ is a prime ideal in $M/I$ and let $AB \subseteq P$. Note that neither $A$ nor $B$ do not necessarily contain $I$, but $A + I$ and $B + I$ do. Furthermore, $A + I \subseteq P$ and $B + I \subseteq P$.

We have

$$[(A + I)/I][(B + I)/I] \subseteq [(A + I)/I] \cap [(B + I)/I] \subseteq P/I.$$ 

Since $P/I$ is prime, $(A + I)/I \subseteq P/I$ or $(B + I)/I \subseteq P/I$. Again using the general lattice isomorphism, we conclude that $A \subseteq A + I \subseteq P$ or $B \subseteq B + I \subseteq P$.

Proof for maximal ideals. Trivial consequence of the general lattice isomorphism.

Proof for radical ideals. Correspondence of radical ideals follows from correspondence of prime ideals since any radical ideal satisfies definition 562 (b) and thus equals the intersection of all prime ideals containing it.

Corollary 652. The two-sided ideal $I$ of the ring $R$ is maximal if and only if the quotient $R/M$ is a simple ring.

In particular, if $R$ is commutative, $R/M$ is a field.

See proposition 664 for the corresponding statement for prime ideals in commutative rings.

Proof. Since $M$ is maximal, only $M$ and $R$ are ideals of $R$ containing $M$. Therefore, by theorem 651 (Quotient ideal lattice theorem), $R/M$ has only two ideals. The converse also follows from the lattice theorem.

Definition 653. We say that an $R$-semimodule is noetherian if any of the following equivalent conditions hold:

(a) Every ascending chain of $R$-sub-semimodules stabilizes. That is,

$$N_1 \subseteq N_2 \subseteq N_3 \cdots$$

implies that there exists an index $k_0$ such that $N_k = N_{k_0}$ for $k > k_0$.

This condition is sometimes abbreviated as ACC (ascending chain condition).

(b) Every nonempty family of $R$-sub-semimodules has a maximal element.

(c) Every $R$-sub-semimodule is finitely generated, i.e. is the linear span of finitely many elements.

Proof of correctness.

Proof of equivalence of 653 (a) and 653 (b). Follows from the equivalences in definition 996 adapted to the opposite of the lattice of $R$-sub-semimodules.

Proof that 653 (b) implies 653 (c). Suppose that every nonempty family of sub-semimodules has a maximal element and let $N$ be a sub-semimodule.

Let $K := \text{span} x_1, \ldots, x_n$ be maximal in the family of all finitely-generated $R$-sub-semimodules. Adding any particular element from $N$ does not change $K$, because otherwise it would not be maximal. Thus, $K = N$. 

318
Proof that 653 (c) implies 653 (a). Suppose that every $R$-sub-semimodule is finitely generated and let

$$N_1 \subseteq N_2 \subseteq N_3 \cdots$$

be a chain of $R$-sub-semimodules.

Suppose that for every positive integer $n$, there exists an $R$-sub-semimodule $N_k$ in this chain with more than $n$ elements. Then the union

$$\bigcup_{k=1}^{\infty} N_k,$$

which by proposition 544 (a) is also an $R$-sub-semimodule, contains infinitely many elements, contradicting our initial assumption.

Therefore, every ascending chain of $R$-sub-semimodules stabilizes. \hfill \Box

**Proposition 654.** Noetherian modules over an arbitrary ring $R$ have the following basic properties:

(a) If $M$ is noetherian, then every $R$-submodule of $M$ also is.

(b) Let $N$ be an $R$-submodule of $M$. Then $M$ is noetherian if and only if both $N$ and their quotient $M/N$ are.

**Proof.**

**Proof of 654 (a).** Trivial.

**Proof of 654 (b).** By theorem 607 (Quotient submodule lattice theorem), every chain of submodules of $M/N$ corresponds to a chain of submodules in $M$. Thus, if $M$ is noetherian, clearly $M/N$ also is.

Conversely, suppose that both $N$ and $M/N$ are noetherian. Let

$$K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots$$

be an ascending chain of $R$-submodule of $M$. Then

$$K_1 \cap N \subseteq K_2 \cap N \subseteq K_3 \cap N \subseteq \cdots$$

(222)

is an ascending chain of $R$-submodules of $N$ and

$$(K_1 + N)/N \subseteq (K_2 + N)/N \subseteq (K_3 + N)/N \subseteq \cdots$$

(223)

is an ascending chain of $R$-submodule of $M/N$.

Both (222) and (223) stabilize. Let $n$ be an index such that, for every positive integer $k$, $K_n \cap N = K_{n+k} \cap N$ and $(K_n + N)/N = (K_{n+k} + N)/N$. For a fixed $k$, we will show that $K_n = K_{n+k}$.

Let $x \in K_{n+k}$. If $x \in N$, then $x \in K_n$ since $K_n \cap N = K_{n+k} \cap N$. Suppose that $x \in K_{n+k} \setminus N$. For any $n \in N$, we have $x + n \in K_{n+k} + N$, and hence

$$x + n + N = x + N \in (K_{n+k} + N)/N = (K_n + N)/N.$$
Then there exists some $y \in K_n$ such that $x - y \in N$. Actually
\[ x - y \in K_{n+k} \cap N = K_n \cap N. \]
Since both $y$ and $x - y$ are in $K_n$, so is their sum $x$. Generalizing on $x$, we conclude that
\[ K_n = K_{n+k}. \]
Therefore, the chain (221) stabilizes, implying that $M$ is noetherian.

**Definition 655.** We say that a (not necessarily commutative) semiring is **left noetherian** (resp. right noetherian) if it is a left (resp. right) noetherian semimodule over itself.

Explicitly, any of the following equivalent conditions characterize a left noetherian semiring:

(a) Every ascending chain of left ideals stabilizes.
(b) Every nonempty set of left ideals has a maximal element.
(c) Every left ideal is finitely generated.

**Proposition 656.** For a noetherian ring $R$, the free module $R^n$ is a noetherian module.

*Proof.* We will use induction on $n$. The cases $n = 0$ and $n = 1$ are trivial.

Suppose that $R^{n-1}$ is noetherian. We can identify $R$ with the submodule of $R^n$ generated by the vector $(0, ..., 0, 1)$. Two vectors $\{x_k\}_{k=1}^n$ and $\{y_k\}_{k=1}^n$ in $R^n$ belong to this submodule if and only if $x_k = y_k$ for $k = 1, ..., n-1$.

By proposition 580, these vectors get mapped to the same vector in the quotient $R^n/R$. Then $R^n/R \cong R^{n-1}$, which is noetherian by the inductive hypothesis. By proposition 654 (b), $R^{n-1}$ is noetherian if and only if $R^n$ is noetherian.

Therefore, $R^n$ is noetherian.

**Lemma 657.** Every surjective endomorphism $f : M \to M$ of a noetherian $R$-module $M$ is an isomorphism.

*Proof.* Consider the equation
\[ f(f(x)) = 0_M. \]
It is obviously satisfied for $x \in \ker f$, but it is also possible that $f(x) \neq 0_M$ while $f(f(x)) = 0_M$. Therefore,
\[ \ker f \subseteq \ker f^2 \subseteq \ker f^3 \subseteq \cdots, \]
where $f^k$ is $k$-fold iterated composition.

Since $M$ is noetherian, this chain stabilizes. Suppose that $\ker f^n = \ker f^{n+k}$ for every positive integer $k$.

Let $y \in \ker f^n$. Since $f$ is surjective, so is $f^n$, and hence there exists some $x$ be such that $f^n(x) = y$. Then $f^n(y) = f^n(f^n(x)) = 0_M$. But ker $f^n = \ker f^{2n}$, hence $x \in \ker f^n$. Therefore, $y = f^n(x) = 0$.

It follows that $f^n$ has a trivial kernel. Then so does $f$. By proposition 457 (g), this implies that $f$ is injective, and hence an isomorphism.
Proposition 658. Consider the free module $R^n$ for a noetherian ring $R$. If the endomorphism $\varphi : R^n \to R^n$ is surjective, then it is also injective and hence an automorphism.

Proof. Follows from proposition 656 and lemma 657.

Theorem 659 (Hilbert’s basis theorem). If $R$ is a noetherian commutative ring, then so is $R[X]$.

Proof. Let $I \subseteq R[X]$ be an arbitrary ideal. We will prove that $I$ is finitely generated.

Denote by $L$ the set of all leading coefficients of polynomials in $I$. The leading coefficient of the product $p(X)q(X)$ of univariate polynomials is the product of their leading coefficients, hence $L$ is an ideal as a consequence of $I$ being an ideal.

As a consequence of $R$ being noetherian, $L$ is finitely generated. Suppose that $L = \{l_1, \ldots, l_n\}$.

For every generator $l_k$, there exists a polynomial $p_k(X)$ in $I$ whose leading coefficient is $l_k$. Denote by $d_k$ the degree of $p_k$ and let $d$ be the maximum of the degrees. We will show that $I$ itself is equal to the sum of the finitely generated ideals

$J := \langle p_1, \ldots, p_n \rangle + \langle X, X^2, \ldots, X^d \rangle.$

Let $f(X)$ be some polynomial from $I$ whose leading term is $lX^m$.

We proceed by induction on $m$ to show that $f(X)$ belongs to $J$.

1. If $m \leq d$, then $f(X)$ belongs to the second ideal $\langle X, X^2, \ldots, X^d \rangle$.
2. Suppose that $m > d$ and that every polynomial in $I$ of degree less than $m$ belongs to $J$.

Since $l \in L$, it is a linear combination $l = \sum_{k=1}^{n} t_k l_k$ with coefficients in $R$. Consider the polynomial

$p(X) := \sum_{k=1}^{n} t_k p_k(X)X^{m-d_k}.$

Define $r(X) := f(X) - p(X)$. Since $p(X)$ belongs to $I$, $r(X)$ does too. It is a polynomial in $I$ of degree less than $m$, hence it belongs to $J$. Then

$f(X) = p(X) + \langle r(X) \rangle.$

Hence, $f(X) \in J$.

Our choice of polynomial $f(X) \in I$ was arbitrary. Therefore,

$I \subseteq \langle p_1, \ldots, p_n \rangle + \langle X, X^2, \ldots, X^d \rangle \subseteq I,$

demonstrating that $I$ is finitely generated.
**Definition 660.** Let $M$ be an $R$-algebra over a commutative ring $R$ and fix a subset $B \subseteq M$. We say that the elements of $B$ are **algebraically independent** if any of the following conditions hold:

(a) If $\{u_b\}_{b \in B}$ is a root of some polynomial with indeterminates $\{X_b\}_{b \in B}$ and coefficients in $R$, then it is the zero polynomial.

(b) The evaluation map $\Phi_B : R[X_b \mid b \in B] \to M$ is injective.

Unsurprisingly, if the elements of $B$ are not algebraically independent, we say that they are **algebraically dependent**.

Compare this concept to linear dependence.

**Proof of correctness.**

**Proof that 660 (a) implies 660 (b).** Suppose that $\Phi_B$ is injective and that there exists a polynomial $p(\{X_b\}_{b \in B})$ such that $\Phi_B(p) = 0_M$. Then for any other polynomial $q(\{X_x\}_{b \in B})$, we have $\Phi_B(pq) = 0_M$, and hence either $p$ is the zero polynomial or the evaluation map is not injective.

**Proof that 660 (b) implies 660 (a).** Conversely, suppose that $B$ is a root only of the zero polynomial. Let $\Phi_B(p) = \Phi_B(q)$. Then $B$ is a root of $p - q$ and hence the latter is the zero polynomial. But this implies that $p = q$. Hence, the evaluation map is injective.

**Proposition 661.** Algebraic (in)dependence for the integral domain $D$ has the following basic properties:

(a) Monomials for different indeterminates are algebraically independent over $D$.

(b) Every two univariate polynomials in $D$ are algebraically dependent over $D$.

(c) Every $n + 1$ polynomials in $D[X_1, \ldots, X_n]$ are algebraically dependent over $D$.

**Proof.**

**Proof of 661 (a).** Suppose that the monomials $X_1, \ldots, X_n$ are algebraically dependent over $D$. Then there exists some nonzero polynomial $f(Y_1, \ldots, Y_n)$ such that the evaluation $\Phi_{X_1,\ldots,X_n}(f)$ is the zero polynomial. But the evaluation simply renames the variables, hence $f$ itself is zero. But we have assumed that it is nonzero.

The obtained contradiction demonstrates that $X_1,\ldots,X_n$ are algebraically independent over $D$.

**Proof of 661 (b).** Fix polynomials $p(X)$ and $q(X)$ over $D$. We will construct a polynomial $f(Y, Z)$ over $D$ such that $\Phi_{p,q}(f) = 0$.

If $p(X)$ is zero, simply define $f(Y, Z) := Z$. If $q(X)$ is zero, put $f(Y, Z) := Y$.

Suppose that both are nonzero; denote by $n$ be the degree of $p(X)$ and by $m$ the degree of $q(X)$. We will consider polynomials of the form $p^l q^k$.

Fix a positive integer $d$. We want the degree of $p^l q^k$ to be at most $d$. If

$$l < \frac{d}{2n} \quad \text{and} \quad k < \frac{d}{2m},$$

then $p^l q^k$ is not the zero polynomial, since $p$ and $q$ are both nonzero. Therefore, $\Phi_{p,q}(p^l q^k) = 0_M$. But any polynomial of the form $p^l q^k$ is algebraically dependent over $D$.

322
then, by proposition 637 (b),
\[ \deg(p^l q^k) = nl + km < \frac{d}{2} + \frac{d}{2} = d. \]

These polynomials are all in
\[ L_d := \text{span}\{1, X, X^2, X^3, \ldots, X^{d-1}\}. \]

This is a module of rank \( d \).
Furthermore, there are \( d^2/(4nm) \) such polynomials. If \( d > 4nm \), there are more polynomials of the form \( p^l q^k \) than the rank of \( L_d \). Hence, every \( d + 1 \) such polynomials are linearly dependent, and hence there exists some linear combination
\[ a_1 p^l k_1 + \cdots + a_{d+1} p^{l+1} k_{d+1} = 0. \]

We can thus define the following polynomial in \( D[Y, Z] \):
\[ f(Y, Z) := a_1 Y^{l_1} Z^{k_1} + \cdots + a_{d+1} Y^{l_{d+1}} Z^{k_{d+1}}. \]

Then clearly \( \Phi_{p, q}(f) = 0 \), so \( p \) and \( q \) are algebraically dependent over \( D \).

**Proof of 661 (c).** Let \( p_1, \ldots, p_{n+1} \) be polynomials in \( D[X_1, \ldots, X_{n-1}][X_n] \). By proposition 661 (b), the polynomials \( p_n \) and \( p_{n+1} \) are algebraically dependent over \( D[X_1, \ldots, X_{n-1}] \).

Let \( f(Y_n, Y_{n+1}) \) be a polynomial in \( D[X_1, \ldots, X_{n-1}][Y_n, Y_{n+1}] \) such that \( \Phi_{p_n, p_{n+1}}(f) = 0 \).

The coefficients of \( f \) are themselves polynomials. Let
\[ \hat{f}(Y_1, \ldots, Y_{n-1}, Y_n, Y_{n+1}) \]
be the polynomial obtained from
\[ f(X_1, \ldots, X_{n-1}, Y_n, Y_{n+1}) \]
by renaming the corresponding variables.

Then \( \Phi_{p_1, \ldots, p_{n+1}}(f) = 0 \). Therefore, \( p_1, \ldots, p_{n+1} \) are algebraically dependent over \( D \). \( \square \)
10.9. Integral domains

Definition 662. An integral domain is an entire commutative (unital) ring.

Proposition 663. Integral domains have the following basic properties:

(a) Any subring of an integral domain is also an integral domain.

(b) A commutative ring $R$ is an integral domain if and only if its polynomial ring $R[X]$ is.

(c) In an integral domain, the multiset of roots of a univariate nonzero polynomial of degree $n$ has multiset cardinality at most $n$.

In other words, a polynomial of degree cannot have more that $n$ roots, counting multiple roots.

Proof.

Proof of 663 (a). Trivial.

Proof of 663 (b).

Proof of necessity. If $R[X]$ is an integral domain, by proposition 663 (a), so is $R$.

Proof of sufficiency. It is sufficient to prove the statement for one indeterminate. If $p(X)$ and $q(X)$ are nonzero polynomials, then so is $p(X)q(X)$ by proposition 637 (b).

Proof of 663 (c). We will use induction on the degree. Zero-degree polynomials clearly have zero roots. Suppose that the statement holds for polynomials of degree $n - 1$, and let $p(X)$ have degree $n$.

If $p(X)$ has a root $u$, by the equivalence in definition 641, $(X - u)$ divides $p(X)$. Then $p(X)/(X - u)$ has degree $n - 1$ by proposition 637 (b). Let $M$ be the multiset of roots of $p(X)/(X - u)$. After adding $u$ to $M$, we have incremented its total cardinality by 1, thus making it at most $n$.

Proposition 664. The ideal $P$ of the commutative ring $R$ is prime if and only if the quotient ring $R/P$ is an integral domain.

See corollary 652 for the corresponding statement for maximal ideals in possibly noncommutative rings.

Proof.

Proof of sufficiency. Suppose that $P$ is a prime ideal. Clearly $R/P$ is a commutative ring. Since $P$ is a proper ideal, $R/P$ must be nontrivial. We will show that it is an entire ring.

Let $[x][y] = [0] = P$ (where $[x] = x + P$ is the coset of $x$ in $R/P$). By definition,

$$[x][y] = (x + P)(y + P) = (xy + P),$$

which implies $xy + P = P$ and hence $xy \in P$. Since $P$ is prime, by proposition 558 (h), we have $x \in P$ or $y \in P$.

Therefore, $[x] = [0]$ or $[y] = [0]$. Generalizing on $x$ and $y$, we can conclude that $R/P$ is entire, and thus an integral domain.
Proof of necessity. Suppose that $R/P$ is an integral domain. Since $R/P$ is nontrivial, $P$ must be a proper ideal. We will show that it satisfies proposition 558 (h).

Let $xy \in P$. We have

$$P = [0] = [xy] = [x][y],$$

hence $[x]$ and $[y]$ are zero divisors in $R/P$. But $R/P$ is entire, hence either $[x]$ or $[y]$ must be zero. That is, either $x \in P$ or $y \in P$.

Generalizing on $x$ and $y$, we can conclude that $P$ is a prime ideal. □

Definition 665. We will introduce several notions related to divisibility in integral domains.

(a) We say that $x$ and $y$ are associates if any of the following conditions hold:

(i) Both $x$ | $y$ and $y$ | $x$.

(ii) There exists a unit $u$ such that $x = uy$.

(iii) The principal ideals $\langle x \rangle$ and $\langle y \rangle$ are equal.

(b) We say that the nonzero nonunit element $x$ is irreducible if any of the following conditions hold:

(i) Whenever $x = yz$, then $y$ or $z$ is a unit.

(ii) $\langle x \rangle$ is maximal among all proper principal ideals. Maximality means that, if $\langle x \rangle \subseteq \langle y \rangle$ for some nonzero nonunit $y$, then $\langle x \rangle = \langle y \rangle$.

(c) We say that the nonzero element $x$ is prime if any of the following equivalent conditions hold:

(i) If $x$ | $yz$, then $x$ | $y$ or $x$ | $z$.

(ii) The ideal $\langle x \rangle$ is prime.

This definition is motivated by lemma 26 (Euclid's lemma).

Proof of correctness.

Proof of 665 (a).

Proof that 665 (a i) implies 665 (a ii). If $x$ | $y$ and $y$ | $x$, then there exist $a$ and $b$ such that $x = ay$ and $y = bx$. Hence, $x = abx$. Since we are working in an integral domain, we can cancel $x$ to obtain $ab = 1_R$. Therefore, both $a$ and $b$ are units.

Proof that 665 (a ii) implies 665 (a iii). Suppose that $x = uy$ for some unit $u$. If $z(x)$, then $x = uy$ divides $z$ and hence $y$ also divides $z$, implying that $\langle x \rangle \subseteq \langle y \rangle$. We obtain the converse inclusion by noting that $y = u^{-1}x$.

Proof that 665 (a iii) implies 665 (a i). If $\langle x \rangle = \langle y \rangle$, then, by proposition 558 (c), $x$ | $y$ and $y$ | $x$.

Proof of 665 (b).
Proof that 665 (b i) implies 665 (b ii). Suppose that \( x \) is not a unit and that \( x = yz \) implies that \( y \) or \( z \) is a unit. Since we are working in an integral domain, \( x \) is necessarily nonzero.

Let \( \langle x \rangle \subseteq \langle w \rangle \) for some nonunit \( w \). By proposition 558 (c), \( w \mid x \). Then there exists some element \( a \) such that \( x = aw \). Since \( w \) is not a unit by assumption, \( a \) must be a unit. By the equivalent definitions of associates in a domain, \( \langle x \rangle = \langle w \rangle \).

Proof that 665 (b ii) implies 665 (b i). Suppose that \( \langle x \rangle \) is maximal among nonzero proper principal ideals.

Let \( x = yz \). If, without loss of generality, \( \langle x \rangle \subseteq \langle y \rangle \), then \( \langle x \rangle = \langle y \rangle \) and, again by the equivalent conditions for associates, there exists some unit \( u \) such that \( x = uy \). Cancelling \( y \) in \( uy = yz \), we obtain \( u = z \). Hence, \( z \) is a unit.

Proof of 665 (c). Trivial.

Proposition 666. The notions from definition 665 have the following basic properties:

(a) Every prime element is irreducible.

(b) An element of the domain \( D \) is irreducible in \( D \) if and only if it is irreducible in \( D[X] \).

(c) If \( \varphi : D \rightarrow E \) is an isomorphism, then \( x \) and \( y \) are associates in \( D \) if and only if \( \varphi(x) \) and \( \varphi(y) \) are associates in \( E \).

(d) If \( \varphi : D \rightarrow E \) is an isomorphism, then \( x \) is prime (resp. irreducible) in \( D \) if and only if \( \varphi(x) \) is prime (resp. irreducible) in \( E \).

Proof.

Proof of 666 (a). Let \( x \) be a prime element and suppose that \( x = yz \). Then \( x \) divides \( y \) or \( z \). If, without loss of generality, \( x \) divides \( y \), then \( x \) and \( y \) are associates, and, by the equivalence of conditions in definition 665 (a), \( z \) must be a unit.

Proof of 666 (b).

Proof of sufficiency. Suppose that \( x \) is irreducible in \( D \) and let \( x = y(X)z(X) \). By proposition 637 (b), both \( y(X) \) and \( z(X) \) must be constant polynomials. Therefore, they are scalars, and since \( x \) is irreducible, \( y \) or \( z \) is a unit. By proposition 573 (d), if \( y \) is a unit in \( D \), it is a unit in \( D[X] \).

Generalizing on \( x \), it follows that every irreducible element in \( D \) is also irreducible in \( D[X] \).

Proof of necessity. Suppose that \( x \) is irreducible in \( D[X] \) and let \( x = yz \). Then \( y \) or \( z \) is a unit of \( D[X] \), and thus again by proposition 573 (d), it is a unit of \( D \).

Generalizing on \( x \), it follows that every element of \( D \) that is irreducible in \( D[X] \) is also irreducible in \( D \).

Proof of 666 (c). Follows from proposition 536.

Proof of 666 (d). Follows from proposition 536.

Example 667. We list some examples of the notions from definition 665:
(a) **Prime numbers** are irreducible integers by their definition. By lemma 26 (Euclid’s lemma), they are also prime.

The inverse $-p$ of the prime number $p$ is also irreducible and prime in $\mathbb{Z}$, but convention requires “prime numbers” to be positive.

(b) Consider the ring of univariate polynomials over $\mathbb{R}$ whose constant coefficient is rational.

The polynomial $X$ is irreducible. Indeed, if $X = p(X)q(X)$, by proposition 637 (b), one of $p(X)$ or $q(X)$ must be a constant polynomial, i.e. a unit.

The polynomial $X$ is not prime, however. We have $X \mid (\sqrt{2}X)^2$, but $X \nmid \sqrt{2}X$ because that would imply that $\sqrt{2}$ is a polynomial in our ring, and it is not a rational number.

**Remark 668.** If $x$ and $y$ are associates, we generally have no reason to prefer $x$ to $y$. This leads to a non-uniqueness in certain contexts, e.g. choosing a greatest common divisor or, more generally, a generator for a principal ideal. In such cases, we often prefer working with ideals.

Fortunately, in the majority of cases, we have good candidates for uniqueness:

- In the domain $\mathbb{Z}$ of integers, there are two units, 1 and $-1$. It is conventional to choose the positive greatest common divisor.

- In a polynomial ring over the integers $\mathbb{Z}$, by proposition 573 (d), the units are again 1 and $-1$, and we can choose the leading coefficient to be positive.

- If $\mathbb{K}$ is any field, any polynomial is associated with a unique monic polynomial.

**Remark 669.** The lattice of ideals described in proposition 559 (c) is sometimes too general for our needs. Rather than stating definitions and theorems “up to a multiplication by a unit”, it is often more convenient to state them in terms of principal ideals. For this reason, we sometimes restrict ourselves to a lattice consisting only of principal ideals.

Unfortunately, as demonstrated in example 678 (b), there may not be a least upper bound of principal ideals, and thus set of all principal ideals may fail to be lattice. Whether the principal ideals form a lattice is intimately related to the existence of greatest common divisors. This motivates introducing the concept of greatest common divisor domains defined in definition 670.

The extreme case are principal ideal domains, in which this lattice is the same as the general lattice of ideals.

**Definition 670.** We say that an integral domain is a **greatest common divisor domain** if the subset of all principal ideals forms a lattice. By proposition 673, it is sufficient for only joins or meets to exist.

This definition is discussed in remark 669 and its relation to the usual concept of a greatest common divisor is given in definition 671.
Definition 671. Fix arbitrary elements $x$ and $y$ of a GCD domain.

Their greatest common divisor ideal is the supremum of the principal ideals $(x)$ and $(y)$ in the lattice of principal ideals. As discussed in remark 668, we can often choose a canonical representative from this ideal, which we call “the” greatest common divisor and denote by $\gcd(x, y)$. Even without making a choice, we may denote the ideal itself by $(\gcd(x, y))$.

Dually, the infimum of this lattice is generated by “the” least common multiple $(\mathsf{lcm}(x, y))$.

Due to proposition 674 (a), we can extend GCDs and LCMs to finitely many elements rather than only two.

Remark 672. The greatest common divisor $\gcd(0, 0)$ is often left undefined, but we see no problem with defining it as 0. We have $(0) + (0) = (0)$, and hence the result is consistent with definition 671.

Proposition 673. The greatest common divisors and least common multiples of $x$ and $y$ (have representative that) are related as follows:

$$xy = \gcd(x, y) \mathsf{lcm}(x, y).$$

(224)

Furthermore, if in some integral domain only one of them exists, the other one exists too.

Proof. Let $d$ be any common divisor of $x$ and $y$. Then $d$ also divides the product $xy$. Let $m$, $a$ and $b$ be elements satisfying

$$xy = dm,$$
$$x = da,$$
$$y = db.$$  

We have

$$x(db) = dm.$$ 

Cancelling $d$, we obtain that $x$ divides $m$. We similarly obtain that $y$ divides $m$.

Now suppose that $d$ is the least common divisor. Let $m'$ be a common multiple. We have $xy = m'd'$ for some $d'$. By duality to what we have just proved, $d'$ must be a common divisor. We have $d' | d$, which in turn implies $m | m'$.

Therefore, $m$ is a least common multiple. 

Proposition 674. Greatest common divisor domains have the following basic properties:

(a) A a lattice join, GCD is associative, commutative and idempotent when considered as a binary operation.

(b) If the polynomial ring $R[X]$ over a commutative ring $R$ is a greatest common divisor domain, then $R$ also is. 

The converse to this is true, but it is more difficult to prove. See proposition 691.

(c) Every irreducible element is prime.

Proof.
Proof of 674 (a). Follows from proposition 1255 (a).

Proof of 674 (b). By proposition 663 (a), \( R \) is an integral domain.

Given elements \( x \) and \( y \) from \( D \), \( \gcd(x, y) \) is a polynomial in \( R[X] \) that divides each of the elements. Hence, by proposition 637 (b), the GCD is a constant polynomial. We may thus regard it as an element of \( R \).

Therefore, \( R \) is also a greatest common divisor domain.

Proof of 674 (c). Suppose that \( x \) is an irreducible element and let \( x | yz \). If \( x \) is a unit, it divides both \( y \) and \( z \), and hence it is prime.

Suppose that \( x \) is not a unit. Putting \( d := \gcd(y, z) \), we obtain

\[
x | xy = d \cdot \frac{y}{d} \cdot \frac{z}{d},
\]

where \( d, \frac{y}{d} \) and \( \frac{z}{d} \) are coprime. Hence, \( x \) divides exactly one of the three.

By transitivity of divisibility, \( x \) necessarily divides \( y, z \) or both. \( \square \)

Proposition 675. In a GCD domain, if \( d \) divides both \( x \) and \( y \), and if there exist elements \( a \) and \( b \) such that \( ax + by = d \), then \( d \) is a greatest common divisor of \( x \) and \( y \).

Proof. Suppose that \( e \) is a common divisor of \( x \) and \( y \). Then \( e \) divides both \( ax \) and \( by \), hence also \( ax + by = d \). Therefore, \( e \) divides \( d \), implying that \( d \) is the greatest common divisor of \( x \) and \( y \). \( \square \)

Definition 676. We say that the nonzero elements \( x \) and \( y \) in a GCD domain are coprime if any the greatest common divisor of \( x \) and \( y \) is a unit.

In this form, the definition is unfortunately inconsistent with coprime ideals. See proposition 683 (e).

Definition 677. An irreducible factorization or simply factorization of a nonzero element \( x \) in an arbitrary integral domain is a finite sequence \( p_1, \ldots, p_n \) of irreducible elements such that, for some unit \( u \),

\[
x = up_1 \cdots p_n.
\]

By cancellation of multiplication, the unit \( u \) is uniquely determined by the irreducible factors. If \( x \) is itself a unit, then it is its own factorization (i.e. \( n = 0 \)) because \( p | x \) for an irreducible \( p \) would imply that \( p \) is a unit and hence not irreducible.

We say that two factorizations

\[
x = up_1 \cdots p_n = vq_1 \cdots q_m
\]

are equivalent if \( n = m \) and if there exists a permutation \( \pi \in S_n \) such that \( q_{\pi(k)} \) and \( p_k \) are associated for every \( k = 1, \ldots, n \).

Finally, if any two factorizations of \( x \) are unique, we say that \( x \) factors uniquely into a product of irreducible factors.

Example 678. We list some examples of irreducible factorization:
(a) By definition 677, every integer has a unique factorization.

(b) Consider the ring of real polynomials with a rational constant term discussed in example 667 (b).

We have

\[ 2X^2 = (\sqrt{2}X) \cdot (\sqrt{2}X) = 2 \cdot X \cdot X. \]

We have shown in example 667 (b) that \( X \) is irreducible, and we can similarly show that \( \sqrt{2}X \) is irreducible. Thus, the above are two irreducible factorizations. They are not equivalent, however, since \( X \nmid \sqrt{2}X \).

As a consequence, both \( X \) and \( \sqrt{2}X \) are maximal divisors of \( 2X^2 \), but they are not associates.

**Proposition 679.** Irreducible factorizations in integral domains have the following basic properties:

(a) If every ascending chain of principal ideals stabilizes in the sense of definition 653 (a), then every element has at least one irreducible factorization.

This condition is sometimes abbreviated as ACCP (ascending chain condition on principal ideals).

(b) If every irreducible element is prime, then all factorizations of an element are equivalent (but there may be elements without a factorization).

(c) For any domain \( D \), \( x = u p_1 \cdots p_n \) is an irreducible factorization in \( D \) if and only if it is an irreducible factorization of \( x \) in \( D[X] \).

**Proof.**

**Proof of 679 (a).** Suppose that every ascending chain of principal ideals stabilizes.

Let \( x \) be an arbitrary element, and suppose that it does not have an irreducible factorization. In particular, \( x \) is not a unit and is not irreducible. Then there exist elements \( a_1 \) and \( b_1 \), which are not both irreducible and not both units, such that \( x = a_1 b_1 \). Via natural number recursion, we can build a sequence \( a_1, a_2, \ldots \) such that

\[ \cdots | a_2 | a_1 | x \]

and no two elements are associates. By proposition 558 (c), this implies

\[ \langle x \rangle \nsubseteq \langle a_1 \rangle \nsubseteq \langle a_2 \rangle \cdots . \]

The existence of such a sequence contradicts the ACCP. Therefore, \( x \) must have at least one irreducible factorization.
Proof of 679 (b). Suppose that every irreducible element is prime.

We will prove by induction that, if \( x \) has a factorization of length \( n \), then any other factorization is equivalent to it.

If \( n = 0 \), then \( x = u \) is a unit, and hence \( m = 0 \) and \( x = u = v \).

Otherwise, suppose that factorizations of length \( n - 1 \) are unique and that we are given the factorizations

\[
x = up_1 \cdots p_n = vq_1 \cdots q_m.
\]

(225)

Since \( p_1 \) is prime, \( p_1 \) divides \( q_{k_0} \) for some \( k_0 = 1, \ldots, m \). Thus, \( p_1 = wq_k \) for some unit \( w \).

We can thus cancel \( p_1 \) to obtain

\[
up_2 \cdots p_n = (uw)q_1 \cdots q_{k_0-1}q_{k_0+1} \cdots q_m.
\]

By the inductive hypothesis, this factorization is unique. Hence, \( n = m \), and there exists a permutation \( \pi \in S_n \) such that \( p_k = q_{\pi(k)} \) for \( k = 2, \ldots, n \).

Then

\[
\hat{\pi}(k) := \begin{cases} k_0, & k = 1 \\ \pi(i), & k > 1 \end{cases}
\]

is a permutation witnessing the equivalence of the factorizations (225).

Proof of 679 (c).

Proof of sufficiency. Suppose that

\[
x = u(X)p_1(X) \cdots p_n(X)
\]

(226)

be an irreducible factorization of \( x \in D \) in \( D[X] \). By proposition 637 (b), all polynomials in this factorization are constants. By proposition 666 (b), since they are irreducible in \( D[X] \), they are also irreducible in \( D \).

Therefore, (226) is an irreducible factorization of \( x \) in \( D \).

Proof of necessity. Due to proposition 573 (d) and proposition 666 (b), irreducible elements and units in \( D \) are also irreducibles and units in \( D[X] \). Hence, every irreducible factorization in \( D \) is also an irreducible factorization in \( D[X] \).

Definition 680. We say that an integral domain is a unique factorization domain if any of the following equivalent conditions hold:

(a) Every element factors uniquely into a product of irreducible elements.

(b) Every ascending chain of principal ideals stabilizes and every irreducible element is prime.

Proof of correctness.

Proof that 680 (a) implies 680 (b).
**Proof that irreducibles are prime.** holds, hence we only need to prove that an irreducible element is prime.

Let \( x \) be an irreducible element. Suppose that \( x \mid yz \), so that there exists some element \( a \) satisfying \( ax = yz \). Let \( y = u_1 \cdots u_n \) and \( z = v_1 \cdots v_m \) be irreducible factorizations. Then

\[
ax = (uv)p_1 \cdots p_n q_1 \cdots q_m.
\]

The factorizations are equivalent, hence \( x \) must divide one of the other irreducible elements. If \( x \mid p_k \) for some \( k = 1, \ldots, n \), then \( x \mid y \). If \( x \mid q_k \) for some \( k = 1, \ldots, m \), then \( x \mid z \).

**Proof of ascending chain condition.** Suppose that every element has a unique factorization.

Fix an ascending sequence of principal ideals

\[
\langle x_1 \rangle \subseteq \langle x_2 \rangle \subseteq \langle x_3 \rangle \cdots.
\]

By proposition 558 (c), this implies that \( x_{k+1} \) divides \( x_k \) for \( k = 1, 2, \ldots \)

Let \( x_1 = yx_2 \) and fix irreducible factorizations

\[
x_1 = up_1 \cdots p_n
\]
\[
x_2 = vq_1 \cdots q_m
\]
\[
y = wr_1 \cdots r_k.
\]

Since the factorizations are unique, we have \( n = m + k \). If \( k = 0 \), then \( x_1 \) and \( x_2 \) are associated and \( \langle x_1 \rangle = \langle x_2 \rangle \). If \( k > 0 \), then \( \langle x_1 \rangle \subsetneq \langle x_2 \rangle \), and \( x_2 \) has a strictly shorter irreducible factorization.

Proceeding by induction on the length of the factorization, we conclude that there are at most \( n \) strict inclusions in the sequence of ideals.

**Proof that 680 (b) implies 680 (a).** Follows from proposition 679 (a) and proposition 679 (b).

**Proposition 681.** Unique factorization domains have the following basic properties:

(a) If the polynomial ring \( R[X] \) over a commutative ring \( R \) is a unique factorization domain, then \( R \) also is.

The converse to this is true, but it is more difficult to prove. See proposition 695.

(b) Every unique factorization domain is a GCD domain.

Proof.

**Proof of 681 (a).** Suppose that \( D[X] \) is a unique factorization. By proposition 679 (c), every irreducible factorization of \( x \in D \) in \( D[X] \) is also an irreducible factorization in \( D[X] \). This implies both existence and uniqueness.

**Proof of 681 (b).** Fix arbitrary elements \( x \) and \( y \). We will show that they have a greatest common divisor.
Suppose that we have the decompositions
\[ x = u_1 \cdots u_n \]
\[ y = v_1 \cdots v_m. \]

Otherwise, let \( r_1, \ldots, r_k \) be a well-ordering on the set \( \{p_1, \ldots, p_n, q_1, \ldots, q_m\} \). For every \( i = 1, \ldots, k \), let \( s_i \) be the minimum of the number of elements from \( p_1, \ldots, p_n \) associated with \( r_i \) and the number of elements of \( q_1, \ldots, q_m \) associated with \( r_i \). Finally, define
\[ r := r_1^{s_1} \cdots r_k^{s_k}. \]

We can use nested induction on \( k \) and \( s_k \) to show that \( r \) is, up to a unit, the greatest common divisor of \( x \) and \( y \). Indeed, \( k = 0 \) implies that \( x \) and \( y \) have no common divisors. For the inductive step, note that dividing both \( x \) and \( y \) by \( r_k^{s_k} \) allows us to use the inductive hypothesis with \( r_1^{s_1} \cdots r_{k-1}^{s_{k-1}} \), and that multiplying back by \( r_k^{s_k} \) makes \( r \) the GCD of \( x \) and \( y \).

**Definition 682.** We say that an integral domain is a **principal ideal domain** if every ideal is principal.

**Proposition 683.** Principal ideal domains have the following basic properties:

(a) Every principal ideal domain is noetherian.

(b) Every principal ideal domain is a unique factorization domain.

Example 686 demonstrates that the converse is not true.

(c) Prime ideals in a principal ideal domains are maximal.

(d) For the greatest common divisor, we have
\[ \langle \gcd(x, y) \rangle = \langle x \rangle + \langle y \rangle = \langle x, y \rangle \]

and for the least common multiple,
\[ \langle \lcm(x, y) \rangle = \langle x \rangle \cap \langle y \rangle. \]

(e) \( x \) and \( y \) are coprime elements if and only if \( \langle x \rangle \) and \( \langle y \rangle \) are coprime ideals.

**Proof.**

**Proof of 683 (a).** In a principal ideal domain every ideal is generated by a single element, hence the domain satisfies definition 655 (c). Thus, it is noetherian.

**Proof of 683 (b).** Suppose that \( x \) is an irreducible element of a principal ideal domain. Then \( x \) satisfies definition 665 (b ii), and hence \( \langle x \rangle \) is a maximal ideal. By proposition 558 (e), maximal ideals are prime, and hence \( x \) is a prime element.

Therefore, every irreducible element is prime. Combined with proposition 683 (a), this implies that the domain satisfies definition 680 (b), and is hence a unique factorization domain.

333
Proof of 683 (c). Let \( P \) be a prime ideal in a principal ideal domain. Then \( P = \langle p \rangle \) for some prime element \( p \). By proposition 666 (a), \( p \) is irreducible, and hence \( \langle p \rangle \) is a maximal ideal.

Proof of 683 (d). Follows from proposition 559 (c) by nothing that every ideal is principal.

Proof of 683 (e). Follows from proposition 683 (d).

Theorem 684 (Bezout’s identity). In a principal ideal domain, for every pair of elements \( x \) and \( y \), there exist some elements \( a \) and \( b \) such that

\[
ax + by = \gcd(x, y). \tag{227}
\]

Algorithm 690 (Extended Euclidean algorithm) gives us an explicit construction for \( a \) and \( b \). The converse, proposition 675, holds more generally.

Proof. Clearly \( x/ \gcd(x, y) \) and \( y/ \gcd(x, y) \) are coprime elements. By proposition 683 (e), there exist elements \( a \) and \( b \) such that

\[
a\frac{x}{\gcd(x, y)} + b\frac{y}{\gcd(x, y)} = 1.
\]

Multiplying by \( \gcd(x, y) \), we obtain (227).

Corollary 685. The multiplicative group \( \mathbb{Z}_n^\times \) of the ring \( \mathbb{Z}_n \) of integers modulo \( n > 1 \) is the set of all positive integers coprime to \( n \).

In particular, \( \mathbb{Z}_p \) is a field if and only if \( p \) is a prime number.

Proof. Note that \( x < n \) is invertible modulo \( n \) if and only if there exists an integer \( a \) such that \( ax = 1 \pmod{n} \). That is, if there exist integers \( a \) and \( b \) such that \( ax + bn = 1 \).

The rest of the theorem follows from theorem 684 (Bezout’s identity) in one direction and proposition 675 in the other direction.

Example 686. The unique factorization domain \( \mathbb{Z}[X] \) is not a principal ideal domain.

Note that \( \mathbb{Z}[X] \) is a unique factorization domain by proposition 681 (a).

Consider the ideal \( I \) of polynomials with an even constant term. Assume that \( I \) is generated by the polynomial \( p(X) \in \mathbb{Z}[X] \). Since \( 2 \in I \), then \( p(X) \) divides 2, so \( p(X) \in \{-2, -1, 1, 2\} \). But then \( p(X) \) is a unit, and hence, \( I = \langle p(X) \rangle = \mathbb{Z}[X] \), which contradicts the definition of \( I \).

The obtained contradiction proves that \( \mathbb{Z}[X] \) is not a principal ideal domain.

Definition 687. An Euclidean domain is an integral domain \( D \) endowed with a function \( \delta : D \to \mathbb{Z}_{\geq 0} \), which we call the Euclidean degree, such that for every pair \( x \) and \( y \) of elements of \( D \) with \( y \neq 0_D \), there exists a pair \( q \) and \( r \) such that

\[
x = yq + r \quad \tag{228}
\]

holds and either \( r = 0_D \) or \( \delta(r) < \delta(q) \).

We say that \( y \) divides \( x \) with quotient \( q \) and remainder \( r \).

If the quotient and remainder are unique, as they usually are, we use the special notation

\[
\text{quot}(x, y) := q,
\]

\[
\text{rem}(x, y) := r = x - y \text{ quot}(x, y).
\]

334
Proposition 688. Euclidean domains have the following basic properties:

(a) Every Euclidean domain is a principal ideal domain.

Counterexamples to the converse are discussed in [And88].

(b) Every field is a Euclidean domain.

(c) A commutative ring $R$ is a field if and only if its polynomial ring $R[X]$ is a principal ideal domain.

Furthermore, if $R$ is a field, the polynomial degree function $\deg : R[X] \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ makes $R$ an Euclidean domain.

We are free to define $\deg$ to take any value for the zero polynomial.

Proof.

Proof of 688(a). Fix an ideal $I$ of the Euclidean domain $D$. By proposition 11, the set $\delta(I)$ has a minimum. Choose an element $m \in I$ such that $\delta(m) = \min \delta(I)$. Obviously $\langle m \rangle \subseteq I$.

We will prove that $I \subseteq \langle m \rangle$.

Let $x \in I$. We divide it by $m$ to obtain

$$x = mq + r,$$

such that either $r$ is zero or $\delta(r) < \delta(m)$. Since both $x$ and $m$ are in $I$, we have $r = mq - x \in I$.

But $m$ minimizes $\delta$ over $I$, thus $\delta(m) \leq \delta(r)$, which contradicts $\delta(r) < \delta(m)$.

Therefore, $r$ is zero and

$$x = mq,$$

which implies that $x \in \langle m \rangle$. This proves $I \subseteq \langle m \rangle$.

We have now obtained $\langle m \rangle = I$. Since $I$ was an arbitrary ideal, we conclude that every ideal in the domain is principal.

Proof of 688(b). By proposition 598, a field is necessarily an integral domain. Since every element of $\mathbb{K}$ is divisible (without remainder), the Euclidean function can be arbitrary; for definiteness, we take it to be canonically zero.

Proof of 688(c).

Proof of sufficiency. Suppose that $R[X]$ is a principal ideal domain.

By proposition 663 (a), $R$ is an integral domain. By theorem 649 (Quotient algebra universal property), $R[X]/\langle X \rangle \cong R$. By proposition 664, $\langle X \rangle$ is a prime ideal in $R[X]$. By proposition 683 (c), $\langle X \rangle$ is a maximal ideal. By corollary 652, $R[X]/\langle X \rangle \cong R$ is a field.

Proof of necessity. Suppose that $R$ is a field. By proposition 663 (b), $R[X]$ is a domain. By algorithm 638 (Euclidean division of polynomials), we can divide two polynomials $f(X)$ and $g(X)$, where $g(X)$ is monic, in a way that satisfies the definition of an Euclidean domain.

Thus, for any $f(X)$ and a nonzero $g(X)$ with leading coefficients $b_m$, algorithm 638 (Euclidean division of polynomials) gives us

$$f(X) = g(X) \frac{q(X)}{b_m} + r(X).$$

\qed
**Algorithm 689** (Euclidean algorithm). In an Euclidean domain, we can explicitly construct the greatest common divisor of arbitrary elements $x$ and $y$ as follows:

(a) If $y$ is zero, halt the algorithm with $\gcd(x, y) := x$.

(b) Define $r_{-1} := x$ and $r_0 := y$.

(c) Starting with $k = 1$, obtain a quotient $q_k$ and remainder $r_k$ so that

$$r_{k-2} = r_{k-1}q_k + r_k.$$

If $r_k$ is not zero, repeat **algorithm 689** (c) with $k + 1$.

Otherwise, halt the algorithm with $\gcd(x, y) := r_{k-1}$.

*Proof of correctness.* If $y$ is zero, then $\langle x, y \rangle = \langle x \rangle$, and hence the result is consistent with definition 671.

Otherwise, Euclidean division ensures that $\delta(r_k) < \delta(r_{k-1})$ on the $k$-th step. Thus, the algorithm halts. Denote by $n$ the last step where $r_n$ is not zero.

We show by induction on $k < n$ that $r_n$ divides $r_{n-k}$. The base case $k = 0$ is obvious since $r_n$ divides itself.

Assume that $r_n$ divides $r_{n-i}$ for $0 \leq i < k$. Now, since

$$r_{n-k} = r_{n-(k-1)}q_{n-(k-2)} + r_{n-(k-2)}$$

and both of the terms on the right-hand side are multiples of $r_n$, the left-hand side $r_{n-k}$ is also a multiple.

So, we conclude that $r_n$ divides both $r_{n-(n-1)} = r_1 = y$ and $r_{n-n} = r_0 = x$.

Finally, we must show that $r_n$ is the greatest among all common divisors of $x$ and $y$. Let $d$ be a common divisor. Then $d \mid r_0$ and $d \mid r_1$. Suppose that $d \mid r_{k-1}$ and $d \mid r_{k-2}$, for some $k < n$. Then

$$r_{k-2} = r_{k-1}q_k + r_k,$$

implying that $d$ also divides $r_k$. Hence, we obtain $d \mid r_n$. Since our choice of common divisor $d$ was arbitrary, we conclude that $r_n$ is the greatest common divisor.

**Algorithm 690** (Extended Euclidean algorithm). In an Euclidean domain, for $x$ and $y$ we can explicitly construct elements $a$ and $b$ so that theorem 684 (Bezout’s identity) holds, i.e.

$$ax + by = \gcd(x, y).$$

(a) If $y$ is zero, then halt the algorithm with $a = 1$, $b = 0$.

(b) Let $r_{-1}, r_0, r_1, \ldots, r_n$ and $q_1, q_2, \ldots, q_n$ be the sequences of quotients and remainders from **algorithm 689** (Euclidean algorithm). The extended Euclidean algorithm proceeds as follows:

Define

$$a_k := \begin{cases} 1, & k = 1, \\ a_{k-2} - a_{k-1}q_k, & k > 1, \end{cases}$$
and
\[ b_k := \begin{cases} -q_1, & k = 1, \\ b_{k-2} - b_{k-1}q_k, & k > 1. \end{cases} \]

Halt the algorithm with \( a := a_n \) and \( b := b_n \).

Proof of correctness. Suppose that \( y \) is not zero. We will prove with induction on \( k < n \) that
\[ r_k = ax_k + by_k. \]

For the base case \( k = 1 \), we have
\[ r_{-1} = r_0 q_1 + r_1, \]
\[ x = yq_1 + r_1, \]
\[ x - yq_1 = r_1, \]
hence \( r_1 = x + (-q_1)y = a_1 x + b_1 y \).

For \( k > 1 \), we have
\[ r_{k-2} = r_{k-1} q_k + r_k, \]
\[ x a_{k-2} + y b_{k-2} = (x a_{k-1} + y b_{k-1}) q_k + r_k, \]
\[ x(a_{k-2} - a_{k-1}q_k) + y(b_{k-2} - b_{k-1}q_k) = r_k. \]

This completes the induction.

Finally, since \( r_n = \gcd(x, y) \), we conclude that
\[ \gcd(x, y) = a_n x + b_n y. \]

Proposition 691. If the integral domain \( D \) is a greatest common divisor domain, so is \( D[X] \).

Proof. Let \( D \) be a GCD domain and let \( \mathbb{K} \) be its field of fractions.

Let \( p(X) \) and \( q(X) \) be arbitrary polynomials in \( D[X] \). We will show that they have a greatest common divisor.

By proposition 688 (c), \( \mathbb{K}[X] \) is an Euclidean domain, and thus \( p(X) \) and \( q(X) \) have a GCD in \( \mathbb{K}[X] \), which is unique up to multiplication by a unit in \( \mathbb{K}[X] \). Thus, taking an arbitrary GCD
\[ r(X) = \sum_{k=0}^{n} \frac{a_k}{b_k} X^k, \]
the polynomial \( b_0 \cdots b_n r(X) \) is also a GCD. Furthermore, the latter is actually a polynomial in \( D[X] \).

Therefore, \( p(X) \) and \( q(X) \) have a GCD in \( D[X] \). \( \square \)
**Definition 692.** The content of a univariate polynomial over a GCD domain is the GCD of its coefficients.

Dividing \( p(X) \) by its content, we obtain another polynomial, which we call the **primitive part** of \( p(X) \). Polynomials whose content is a unit are called **primitive** in the context of lemma 693 (Gauss’ lemma), although this clashes with the unrelated concept of primitive polynomials in finite fields.

**Lemma 693** (Gauss’ lemma). If \( p(X) \) and \( q(X) \) are primitive polynomials, then \( p(X)q(X) \) is also primitive.

**Proof.** Fix two primitive polynomials

\[
p(X) = \sum_{k=0}^{n} a_k X^k, \quad q(X) = \sum_{k=0}^{m} b_k X^k.
\]

Let \( d \) be the content of \( p(X)q(X) \). It divides every coefficient

\[
\sum_{i+j=k} a_i b_j
\]

of \( p(X)q(X) \), and hence also \( a_i b_j \) for every particular pair of indices \( i < n \) and \( j < m \).

For any fixed \( i < n \), \( d \) divides \( a_i b_j \) for every \( j < m \). Since \( q(X) \) is primitive, \( d \) cannot divide \( b_j \) for every \( j < m \) unless \( d \) is a unit. Hence, \( d \) necessarily divides \( a_i \). Our choice of \( i \) was arbitrary, hence \( d \) divides \( a_i \) for every \( i < n \). But \( p(X) \) is also primitive. Therefore, \( d \) can only be a unit.

**Lemma 694.** Let \( D \) be a GCD domain and let \( K \) be its field of fractions. If a primitive polynomial \( p(X) \) is irreducible in the polynomial ring \( K[X] \), then it is irreducible in \( D[X] \).

**Proof.** Suppose that \( p(X) \) is primitive and irreducible in \( K[X] \). Let

\[
p(X) = q(X)r(X),
\]

where \( q(X) \) and \( r(X) \) are polynomials from \( D[X] \). Then \( q(X) \) or \( r(X) \) is a unit in \( K[X] \), hence it is a nonzero constant polynomial. Suppose that \( q(X) = q_0 \).

Then \( q_0 \) is an element of \( D \) that divides all coefficients of \( p(X) \). By assumption, the coefficients of \( p(X) \) are coprime. It follows that \( q_0 \) is a unit in \( D \).

**Proposition 695.** If the integral domain \( D \) is a unique factorization domain, so is \( D[X] \).

**Proof.** Let \( D \) be a unique factorization domain and let \( K \) be its field of fractions.

By proposition 691, \( D[X] \) is a GCD domain. Then proposition 674 (c) is satisfied, with by proposition 679 (b) implies that if an element has at least one irreducible factorization, all others are equivalent to it.

We will now show existence of irreducible factorizations.

Let \( p(X) \) be a polynomial in \( D[X] \). By proposition 688 (c), \( p(X) \) has an irreducible factorization

\[
p(X) = uq_1(X) \cdots q_n(X),
\]

338
where \( q_1(X), \ldots, q_n(X) \) belong to \( \mathbb{K}[X] \).

For a fixed index \( i = 1, \ldots, n \), if the polynomial \( q_i(X) \) has the form
\[
q_i(X) = \sum_{k=0}^{n} \frac{a_k}{b_k} X^k,
\]
then \( b_1 \cdots b_n q_i(X) \) is a polynomial in \( \mathbb{D}[X] \). Denote by \( r_i(X) \) the primitive part this polynomial in \( \mathbb{D}[X] \). Since every scalar from \( \mathbb{D} \) is simply a unit in \( \mathbb{K}[X] \), it follows that \( r_i(X) \) is also irreducible in \( \mathbb{K}[X] \).

Hence, for the appropriate scalar \( v \in \mathbb{K} \),
\[
p(X) = vr_1(X) \cdots r_n(X)
\]
is an irreducible factorization in \( \mathbb{K}[X] \).

By lemma 694, \( r_1(X), \ldots, r_n(X) \) are irreducible elements of \( \mathbb{D}[X] \). Thus, if \( v = \frac{c}{d} \) with \( c \) and \( d \) coprime, then
\[
dp(X) = cr_1(X) \cdots r_n(X)
\]
is an irreducible factorization in \( \mathbb{D}[X] \).

By lemma 693 (Gauss’ lemma), the product \( r_1(X) \cdots r_n(X) \) is a primitive polynomial in \( \mathbb{D}[X] \). Hence, \( d \) cannot divide neither \( c \) nor \( r_1(X) \cdots r_n(X) \), implying that \( d \) is a unit. Therefore,
\[
p(X) = cr_1(X) \cdots r_n(X)
\]
is an irreducible factorization of \( p(X) \) in \( \mathbb{D}[X] \). \( \square \)
10.10. Field extensions

**Definition 696.** A **splitting field** for a nonconstant polynomial \( f(X) \in \mathbb{k}[X] \) of degree \( n \) is the smallest field extension \( \mathbb{K} \) of \( \mathbb{k} \) in which \( f(X) \) has \( n \) roots. That is,

\[
\mathbb{K} \cong \mathbb{k}(a_1, \ldots, a_n),
\]

where \( a_1, \ldots, a_n \) are roots of \( f(X) \).

By **proposition 697**, splitting fields exist and are unique up to an isomorphism.

**Proposition 697.** There exists a unique up to an (possibly nonunique) isomorphism splitting field for every nonconstant polynomial in one indeterminate over a field.

**Proof.**

**Proof of existence.** We use induction on the degree of the polynomial \( f(X) = \sum_{k=0}^{n} a_k x^k \) over \( \mathbb{k} \). In the base case \( n = 1 \), \( f(X) \) is already linear, and hence \( \mathbb{k} \) is itself a splitting field for \( f(X) \).

Suppose that there exist splitting fields for polynomials over \( \mathbb{k} \) of degree \( n - 1 \). By **theorem 561** (Maximal ideal theorem), the principal ideal \( (f(X)) \) is contained in some maximal ideal \( M \). By **corollary 652**, the quotient of \( \mathbb{R}[X] \) by \( M \) is a field.

Define \( u_n := (X) + M \). We have \( f(u_n) = (f(X)) + M \), hence \( u_n \) is a root of \( f \) in \( M \). Then

\[
f(X) = (X - u_n)q(X)
\]

for some polynomials \( q(X) \) and \( r(X) \), both of degree less than \( n \) (or \( r(X) = 0 \)).

We can now apply the inductive hypothesis to obtain a splitting field of \( q(X) \). Let \( u_1, \ldots, u_{n-1} \) be the roots of \( q(X) \) in this field. We can then adjoin \( u_1, \ldots, u_n \) to the field \( \mathbb{k} \) to obtain a splitting field \( \mathbb{k}(u_1, \ldots, u_n) \) of \( f(X) \). Denote this field by \( \mathbb{K} \).

**Proof of uniqueness.** Suppose that, given our previous construction, \( \mathbb{L} \) is also a splitting field for \( f(X) \).

Again, we use induction on the degree \( n \) of \( f(X) \). The case \( n = 1 \) is again obvious.

Suppose that any two splitting fields for polynomials over \( \mathbb{k} \) of degree \( n - 1 \) are isomorphic. Let \( b_n \) be a root of \( f(X) \) in \( \mathbb{L} \) and let

\[
f(X) = (X - b_n)r(X).
\]

Let \( b_1, \ldots, b_{n-1} \) be the roots of \( r(X) \). Let \( \varphi \) be an isomorphism between the subfield \( \mathbb{k}(a_1, \ldots, a_{n-1}) \) of \( \mathbb{K} \) and the corresponding subfield \( \mathbb{k}(b_1, \ldots, b_{n-1}) \) of \( \mathbb{L} \). It follows that

\[
\prod_{k=1}^{n-1} (X - b_k) = \prod_{k=1}^{n-1} (X - \varphi(a_k)).
\]

Therefore, we can extend \( \varphi \) to an isomorphism \( \hat{\varphi} : \mathbb{K} \to \mathbb{L} \) by putting \( \hat{\varphi}(a_n) := b_n \). \( \square \)

**Theorem 698** (Classification of finite fields).
(a) The characteristic of a field with \( q \) elements is a prime number \( p \), and \( q \) is a power of \( p \).

The fields of prime cardinality are sometimes called prime fields.

(b) For a prime number \( p \), the ring \( \mathbb{Z}/p \mathbb{Z} \) of integers modulo \( p \) is a field.

(c) All fields with \( q \) elements are isomorphic as splitting fields for the polynomial

\[
X^q - X \in \mathbb{Z}/p \mathbb{Z}[X].
\]

Utilizing the general conventions of identifying isomorphic objects in algebra, we denote by \( \mathbb{F}_q \) "the" finite field with \( q \) elements. Finite fields are also called Galois fields.

Every member of \( \mathbb{F}_q \) is a root of \( X^q - X \).

Proof.

**Proof of 698 (a).** Let \( K \) be a field with \( q \) elements and let \( p \) be the characteristic of \( K \). Then \( \mathbb{Z}/p \mathbb{Z} \) is a subring of \( K \). By corollary 685, \( \mathbb{Z}/p \mathbb{Z} \) is a field.

By corollary 652, \( \langle p \rangle \) is a maximal ideal in \( \mathbb{Z} \), and, by proposition 558 (e), \( p \) is a prime number.

By theorem 468 (Lagrange's theorem for groups), \( p \) divides \( q \). But \( K/\mathbb{Z}/p \mathbb{Z} \) is again a field by theorem 651 (Quotient ideal lattice theorem), and again has prime characteristic. Continuing by induction, we eventually obtain a sequence \( p_1, \ldots, p_n \) of prime numbers such that

\[
q = p_1 \cdots p_n.
\]

By theorem 468 (Lagrange's theorem for groups), \( q \) cannot contain subgroups of prime cardinalities \( p_1 \) and \( p_2 \) unless \( p_1 = p_2 \). Hence, again by induction, we conclude that

\[
p_1 = \cdots = p_n.
\]

Therefore, \( q = p^n \).

**Proof of 698 (b).** Follows from corollary 685.

**Proof of 698 (c).** Let \( K \) be a field with \( q \) elements with characteristic \( p \). We will show that every element of \( K \) is a root of \( X^q - X \in \mathbb{Z}/p \mathbb{Z}[X] \).

The multiplicative group of \( K \) has order \( q - 1 \). The order of a non-zero element \( a \in K \) divides \( q - 1 \) by proposition 483 (a), hence \( a^{q-1} = 1 \pmod{q} \). We also have \( 0^q = 0 \). Therefore, for every element of \( \mathbb{F}_q \), we have \( a^q = a \).

Then

\[
X^q - X = \prod_{u \in K} (X - u).
\]

**Proposition 699.** For every finite field \( \mathbb{F}_q \) and every polynomial ring \( \mathbb{F}_q[X_1, \ldots, X_n] \) in finitely many indeterminates, there exists an \( \mathbb{F}_q \)-algebra isomorphism

\[
\frac{\mathbb{F}_q[X_1, \ldots, X_n]}{\langle X_1^q - X_1 | \ i = 1, \ldots, n \rangle} \cong \text{fun}(\mathbb{F}_q^n, \mathbb{F}_q).
\]

341
where \( \text{fun}(F^n_q, F_q) \) is the \( F_q \)-algebra of all functions from \( F^n_q \) to \( F_q \).

Furthermore, every coset of polynomials has a unique representative given by theorem 233 (Finite field Lagrange interpolation).

Proof. Consider the functional evaluation homomorphism

\[
\Phi : F_q[X_1, ..., X_m] \to \text{fun}(F^m_q, F_q).
\]

By theorem 233 (Finite field Lagrange interpolation), \( \Phi \) is surjective. Then, by theorem 649 (Quotient algebra universal property),

\[
F_q[X_1, ..., X_n]/ \ker \Phi \cong \text{fun}(F^m_q, F_q).
\]

We will now prove that \( \ker \Phi \) equals \( I := \langle X^q_i - X_i \mid i = 1, ..., n \rangle \).

First, let \( e : X \to F_q \) be the variable assignment that assigns \( u_1, ..., u_n \) to the corresponding indeterminates. By theorem 698 (c), every member of \( F_q \) is a root of \( X^q_i - X_i \). Then, for any indeterminate \( X_i \),

\[
\Phi_e(X^q_i - X_i) = u^q_i - u_i \equiv 0 \pmod{q}.
\]

Hence, the polynomial function \( \Phi(X^q_i - X_i) \) is the zero constant function. It follows that any linear combination of the polynomials \( X^q_i - X_i \) for \( i = 1, ..., n \) is also the zero function. Therefore, \( I \subseteq \ker \Phi \).

We will prove the converse inclusion via induction on \( n \).

In the case of a single indeterminate \( X \), for every polynomial \( f(X) \in \ker \Phi \), we know that the entirety of \( F_q \) are roots of \( f(X) \). By proposition 663 (c), \( f(X) \) has at most \( q \) roots. Hence, \( X - u \) divides \( f(X) \) for every \( u \in F_q \). We have

\[
\prod_{u \in F_q} (X - u) \mid f(X),
\]

and hence \( f(X) \in \langle X^q - X \rangle \).

We have, up until now, shown that the entire proposition holds for the case of one indeterminate. Suppose that the proposition holds for \( n - 1 \) indeterminates and let \( f \in F_q[X_1, ..., X_n] \) be a nonconstant polynomial such that \( \Phi(f) \) is the zero function. Due to proposition 573 (b), we can regard \( f \) as a univariate polynomial in \( X_n \) over \( F_q[X_1, ..., X_{n-1}] \). Thus,

\[
f(X_1, ..., X_n) = \sum_{k=0}^{\infty} \left( \sum_{\gamma} a_{(k,\gamma)} X_1^{\gamma_1} X_2^{\gamma_2} ... X_{n-1}^{\gamma_{n-1}} \right) X_n^k,
\]

where \( \gamma \) is a multi-index over the first \( n - 1 \) indeterminates.
As a polynomial in $X_n$, $f$ has $m := (n - 1)p$ roots $s_1, \ldots, s_m$, which are themselves polynomials from $F_q[X_1, \ldots, X_{n-1}]$. For some $c$, we have

$$f(X_1, \ldots, X_n) = c(X_1, \ldots, X_{n-1}) \prod_{j=1}^{m}(X_n - s_j(X_1, \ldots, X_{n-1}))$$

and

$$0 = \Phi(f) = \Phi(c) \cdot \prod_{j=1}^{m}(\Phi(X_n) - \Phi(s_j)).$$

Since $F_q[X_1, \ldots, X_{n-1}]$ is entire, we conclude that either $\Phi(c)$ is the zero function or $\Phi(X_n) = \Phi(s_j)$ for at least one index $1 \leq j \leq m$. The latter is impossible, because $\Phi(X_n)$ is linearly independent from polynomials in the first $n - 1$ variables.

The inductive hypothesis holds for the polynomial $c$, and $\Phi(c)$ being the zero function implies

$$c \in \langle X_i^q - X_i \mid i = 1, \ldots, n - 1 \rangle \subseteq I.$$

Therefore, $f \in I$ since $f$ divides $c$. We have chosen $f$ to be an arbitrary member of $\ker \Phi$, which implies $\ker \Phi \subseteq I$.

We have already shown that $I \subseteq \ker \Phi$. We thus conclude that $I = \ker \Phi$ and

$$F_q[X_1, \ldots, X_m]/I \cong \text{fun}(F_q^m, F_q).$$

\[\square\]

**Definition 700.** We say that the element $a \in K$ of the field extension $K$ of $k$ is **transcendental** over $k$ if it is algebraically independent.

If $a$ is not transcendental, we say that it is **algebraic**. If every element of $K$ is algebraic over $k$, we say that $K$ is an **algebraic extension** of $k$.

**Proposition 701.** Every field is an algebraic extension of itself.

*Proof.* Every element $a \in K$ is a root of the polynomial $X - a$. \[\square\]

**Theorem 702** (Euler’s constant is transcendental). *Euler’s constant $e$ is transcendental over $\mathbb{Q}$.*

**Theorem 703** (Pi is transcendental). *The number $\pi$ is transcendental over $\mathbb{Q}$.*

**Example 704.** *Theorem 703* (Pi is transcendental) implies that the polynomials $\mathbb{Q}[X]$ can be embedded into $\mathbb{R}$ via $\Phi_\pi : \mathbb{Q}[X] \to \mathbb{R}$. We can identify a polynomial

$$p(X) = \sum_{i=0}^{n} a_k X^k$$

with rational coefficients with the number

$$p(\pi) = \sum_{i=0}^{n} a_k \pi^k.$$
**Definition 705.** If \( \mathbb{K} \) is finite-dimensional vector space over \( k \), we say that \( \mathbb{K} \) is a finite extensions of \( k \).

**Lemma 706.** Every finite field extension is algebraic.

**Proof.** Let \( \mathbb{K} \) be a field extension of \( k \). Consider the evaluation map \( \Phi_a : k[X] \to k[u] \) for some \( u \in \mathbb{K} \).

Since the polynomials \( X^k \) for \( k = 0, 1, 2, \ldots \) form a basis for \( k[X] \). If \( \Phi_a \) is injective, then \( \Phi_a(X_k) \) are linearly independent over \( k \). But \( \mathbb{K} \) has finite dimension over \( k \).

The obtained contradiction shows that \( \Phi_a \) is not injective. \( \square \)

**Definition 707.** We say that the field \( \mathbb{K} \) is algebraically closed if any of the equivalent conditions are satisfied:

1. \( \mathbb{K} \) has no nontrivial algebraic extensions.
2. Every irreducible polynomial in \( \mathbb{K}[X] \) is linear.
3. Every nonconstant polynomial in \( \mathbb{K}[X] \) has at least one root in \( \mathbb{K} \).
4. Every polynomial in \( \mathbb{K}[X] \) factors into a product of linear polynomials.
5. Every polynomial in \( \mathbb{K}[X] \) of degree \( n \) has exactly \( n \) roots in \( \mathbb{K} \), counting the root multiplicities.

**Proof.**

**Proof that 707 (a) implies 707 (b).** Let \( p(X) \) be an irreducible polynomial in \( \mathbb{K}[X] \).

Since \( \mathbb{K}[X] \) is a unique factorization domain, it satisfies definition 680 (b), and hence \( p(X) \) is a prime element. Thus, \( \langle p(X) \rangle \) is a prime ideal in \( \mathbb{K}[X] \).

Since \( \mathbb{K}[X] \) is a principal ideal domain, by **proposition 683 (c)**, \( \langle p(X) \rangle \) is also a maximal ideal. By ??, the quotient \( Q := \mathbb{K}[X]/\langle p(X) \rangle \) is a field. The vectors \( 1, X, X^2, \ldots, X^n \) for a basis of \( Q \) over \( k \), where \( n \) is the degree of \( p(X) \).

By **lemma 706**, \( Q \) is an algebraic extension of \( \mathbb{K} \). Since \( \mathbb{K} \) has no nontrivial algebraic extensions, it follows that \( \mathbb{K} = Q \). Thus, \( Q \) has dimension 1, and we have already discussed that \( \dim Q = \deg p \). Therefore, \( p \) is a linear polynomial.

**Proof that 707 (b) implies 707 (c).** Suppose that every irreducible polynomial is linear.

By **proposition 681 (a)**, \( \mathbb{K}[X] \) is a unique factorization domain, and thus there exist irreducible polynomials \( q_1(X), \ldots, q_n(X) \) and a unit \( a \) such that

\[
p(X) = aq_1(X) \cdots q_n(X).
\]

By assumption, the irreducible polynomials are linear, and hence have roots. Therefore, \( p(X) \) has at least one root.

**Proof that 707 (c) implies 707 (d).** Suppose that \( u_1 \) is a root of \( p(X) \). Then \( p(X) \) is divisible by \( (X - u_1) \). Using induction on the degree of \( p(X) \), we can factor \( p(X) \) into

\[
p(X) = a(X - u_1)(X - u_2) \cdots (X - u_n),
\]

where \( a \) is a unit of \( \mathbb{K} \). This is the desired factorization.
Proof that 707 (d) implies 707 (e). Follows from the equivalence in definition 641 by induction on the polynomial degree. By proposition 663 (c), the number of roots is bounded by $n$.

Proof that 707 (e) implies 707 (a). Suppose that every nonconstant polynomial of degree $n$ has exactly $n$ roots in $k$ and let $K$ be an algebraic extension of $k$.

By proposition 663 (c), every polynomial in $K[X]$ has at most $n$ roots. By assumption, every root of every polynomial is contained in $k$. Since $K$ is algebraic over $k$, it follows that every element of $K$ is a root of some polynomial. Therefore, $K = k$.

Proposition 708. An algebraically closed field has no nontrivial finite extension fields.

Proof. Follows from lemma 706 applied to definition 707 (a).

Theorem 709 (Weak nullstellensatz<sup>6</sup>). Let $K$ be an algebraically closed field and let $K[X_1, ..., X_n]$ be its polynomial ring in $n$ indeterminates.

The ideal $M$ of $K[X_1, ..., X_n]$ is maximal if and only if there exist elements $u_1, ..., u_n$ of $K$ such that

$$M = (X_1 - u_1, ..., X_n - u_n).$$

<sup>6</sup>en: zero locus theorem
11. Linear algebra

Linear algebra is a branch of mathematics that is both very accessible and enormously useful. It studies vector spaces, mostly over real or complex numbers, and linear maps between them. For finite-dimensional vector spaces, this reduces to studying matrices, which also leads to a very rich computational theory.

We have discussed the basics of vector spaces in section 10.6 (Modules) in the context of general modules over rings. The key takeaways are:

- Vector spaces, defined incrementally in definition 543, definition 604 and definition 624, with some properties listed in proposition 625.
- Linear combinations, defined in definition 551 and characterized by theorem 552 (Free semimodule universal property), with important remarks in remark 553.
- Linear spans, defined in definition 543 (e) and characterized via proposition 554.
- Linear maps, defined in definition 543 (d), and multilinear maps, defined in definition 564.
- Quotient spaces, discussed in definition 604 (h) for general modules.
- Theorem 606 (Quotient module universal property) and Theorem 607 (Quotient submodule lattice theorem).
- Linear (in)dependence, defined in definition 610 with some properties listed in proposition 612.
- Hames bases, defined in definition 614 with some properties listed in proposition 616.
- Basis decomposition defined in definition 621.
- Vector space dimensions, whose existence and uniqueness is shown in proposition 627.
- Theorem 634 (Rank-nullity theorem).
11.1. Matrices over rings

We will define and prove some fundamental notions about matrices. We will start with matrices over plain sets and end up with matrices over nontrivial noetherian commutative rings. This is about as general as we want to go without the underlying ring being a field. The main benefit is being able to work with the ring of integers or more general semirings, like the tropical semirings.

**Definition 7.10.** Let $S$ be any nonempty set and $n_1, \ldots, n_k$ be positive integers. An array of shape $n_1 \times \cdots \times n_k$ is a function with signature

$$A : \{1, 2, \ldots, n_1\} \times \cdots \times \{1, 2, \ldots, n_k\} \to S.$$  

“Multi-dimensional array” is also used as a term, but we will avoid it because the terminology conflicts with vector space dimensions.

We can regard “$n_1 \times \cdots \times n_k$” simultaneously as a convenient notation and as a Cartesian product of finite ordinals (modulo the fact that finite ordinals are zero-based). In particular:

(a) A two-dimensional array of shape $m \times n$ is usually called a matrix. Let $A$ be an $m \times n$-matrix. We will denote $A$ as

$$A = \{a_{i,j}\}_{i,j=1}^{m,n}$$

or graphically as the table

$$
\begin{pmatrix}
a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m,1} & a_{m,2} & \cdots & a_{m,n}
\end{pmatrix}
$$

(b) A square matrix of order $n$ is simply an $n \times n$ matrix.

(c) A column vector of dimension $m$ is simply a $m \times 1$ matrix

$$
\begin{pmatrix}
a_{1,1} \\
\vdots \\
a_{m,1}
\end{pmatrix}
$$

When $S$ is a semiring $R$, we often identify the set of all $m$-dimensional column vectors with the free semimodule $R^m$.

(d) A row vector of dimension $n$ is simply an $1 \times n$ matrix

$$
\begin{pmatrix}
a_{1,1} & \cdots & a_{1,n}
\end{pmatrix}
$$

**Remark 7.11.** In practice, the terms “vector”, “tuple” and “finite sequence” are used interchangeably. Formally, the concepts differ slightly:
“Vector” refers to an element of a vector space or, more generally, a semimodule. Column vectors and row vectors are important special cases.

Tuples are defined and discussed in definition 950 (c). Tuples are technically indexed families and the latter are defined without reference to functions, which we use to define both arrays and sequences.

Sequences are defined and discussed in definition 951. Formally, a finite sequence of length $n$ is the same as an array of shape $n$.

**Definition 712.** A block matrix is a “matrix of matrices”. That is, a matrix of the form

$$
\begin{pmatrix}
A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\
A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m,1} & A_{m,2} & \cdots & A_{m,n}
\end{pmatrix},
$$

where all $A_{i,j}$ are matrices of compatible dimensions.

We can write the block matrix

$$
\begin{pmatrix}
A & \cdots & B \\
\vdots & \ddots & \vdots \\
C & \cdots & D
\end{pmatrix}
$$

as

$$
\begin{pmatrix}
a_{1,1} & \cdots & a_{1,n} \\
\vdots & \ddots & \vdots \\
a_{m,1} & \cdots & a_{m,n} \\
\vdots & \ddots & \vdots \\
c_{1,1} & \cdots & c_{1,n} \\
\vdots & \ddots & \vdots \\
c_{m,1} & \cdots & c_{m,n}
\end{pmatrix}
\begin{pmatrix}
b_{1,1} & \cdots & b_{1,n} \\
\vdots & \ddots & \vdots \\
b_{m,1} & \cdots & b_{m,n} \\
\vdots & \ddots & \vdots \\
d_{1,1} & \cdots & d_{1,n} \\
\vdots & \ddots & \vdots \\
d_{m,1} & \cdots & d_{m,n}
\end{pmatrix}.
$$

Given any matrix $A = \{a_{i,j}\}_{i,j=1}^{m,n}$, we can represent it via its block matrix of rows

$$
\begin{pmatrix}
a_{1,-} \\
a_{2,-} \\
\vdots \\
a_{n,-}
\end{pmatrix},
$$

and its block matrix of columns

$$
( a_{-,,1} \mid a_{-,,2} \mid \cdots \mid a_{-,,n} )
$$

**Definition 713.** The main diagonal of the matrix $A = \{a_{i,j}\}_{i,j=1}^{m,n}$ is the sequence $a_{1,1}, \ldots, a_{i,j}, \ldots, a_{k,k}$, where $k := \min\{m, n\}$. The antidiagonal is instead $a_{1,k}, \ldots, a_{i,k-j}, \ldots, a_{k,n-k}$. These can be
visualized as follows:

\[
\begin{pmatrix}
  a_{1,1} & a_{1,2} & \cdots & a_{1,k-1} & a_{k,k} & \cdots \\
  a_{2,1} & a_{2,2} & \cdots & a_{2,k-1} & a_{2,k} & \cdots \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\
  a_{k-1,1} & a_{k-1,2} & \cdots & a_{k-1,k-1} & a_{k-1,k} & \cdots \\
  a_{k,1} & a_{k,2} & \cdots & a_{k,k-1} & a_{k,k} & \cdots \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \ddots
\end{pmatrix}
\]

Over a semiring, we say that a square matrix diagonal if all entries outside the main diagonal are zero. For brevity, we write

\[
\text{diag}(a_1, \ldots, a_n) := \begin{pmatrix}
  a_1 & 0 & \cdots & 0 \\
  0 & a_2 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & a_n
\end{pmatrix}
\]

The notation \(\text{diag}(A)\) is also used to denote the sequence of diagonal entries of \(A\).

**Proposition 714.** Denote by \(R_{m \times n}\) the set of \(m \times n\) matrices over the semiring \(R\). We define three operations on matrices:

(a) We define **matrix addition** \(+\) : \(R_{m \times n} \times R_{m \times n} \rightarrow R_{m \times n}\) componentwise as

\[
\begin{pmatrix}
  a_{1,1} & \cdots & a_{1,n} \\
  \vdots & \ddots & \vdots \\
  a_{m,1} & \cdots & a_{m,n}
\end{pmatrix}
+ \begin{pmatrix}
  b_{1,1} & \cdots & b_{1,n} \\
  \vdots & \ddots & \vdots \\
  b_{m,1} & \cdots & b_{m,n}
\end{pmatrix}
= \begin{pmatrix}
  a_{1,1} + b_{1,1} & \cdots & a_{1,n} + b_{1,n} \\
  \vdots & \ddots & \vdots \\
  a_{m,1} + b_{m,1} & \cdots & a_{m,n} + b_{m,n}
\end{pmatrix}
\]

With addition, \(R_{m \times n}\) becomes an commutative monoid with neutral element the zero matrix

\[
\begin{pmatrix}
  0 & 0 & \cdots & 0 \\
  0 & 0 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & 0
\end{pmatrix}
\]

(229)

(b) We define **scalar multiplication** \(\cdot\) : \(R \times R_{m \times n} \rightarrow R_{m \times n}\) as

\[
\lambda \cdot \begin{pmatrix}
  a_{1,1} & \cdots & a_{1,n} \\
  \vdots & \ddots & \vdots \\
  a_{m,1} & \cdots & a_{m,n}
\end{pmatrix} := \begin{pmatrix}
  \lambda a_{1,1} & \cdots & \lambda a_{1,n} \\
  \vdots & \ddots & \vdots \\
  \lambda a_{m,1} & \cdots & \lambda a_{m,n}
\end{pmatrix}
\]

Under addition and scalar multiplication, \(R_{m \times n}\) becomes an \(R\)-semimodule.

(c) We define **matrix multiplication** in two steps. The definition is justified by proposition 717. First, if \(\{a_i\}_{j=1}^n\) is a row vector and \(\{b_j\}_{i=1}^m\) is a column vector, we define their **inner product** as

\[
a \cdot b := \sum_{i=1}^n a_i b_i.
\]
We can now define matrix multiplication \( \cdot : R^{m \times k} \times R^{k \times n} \rightarrow R^{m \times n} \) as

\[
\begin{pmatrix}
a_{1,-} \\
a_{2,-} \\
\vdots \\
a_{m,-}
\end{pmatrix}
\cdot
\begin{pmatrix}
b_{-1} & | & b_{-2} & | & \cdots & | & b_{-n}
\end{pmatrix}
:=
\begin{pmatrix}
a_{1,-} \cdot b_{-1} & a_{1,-} \cdot b_{-2} & \cdots & a_{1,-} \cdot b_{-n} \\
a_{2,-} \cdot b_{-1} & a_{2,-} \cdot b_{-2} & \cdots & a_{2,-} \cdot b_{-n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m,-} \cdot b_{-1} & a_{m,-} \cdot b_{-2} & \cdots & a_{m,-} \cdot b_{-n}
\end{pmatrix}.
\]

If \( n \) and \( m \) are equal, \( R^{n \times n} \) becomes an \( R \)-algebra under matrix multiplication with multiplicative identity the identity matrix of order \( n \)

\[
\text{diag}(1, \ldots, 1) = 
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 1
\end{pmatrix}.
\] (230)

Proof: The semimodule structure is inherited by the semimodule structure on \( R \). We will show that, if \( n = n \), matrix multiplication is associative and bilinear. Fix matrices

\[
A = \{a_{i,j}\}_{i,j=1}^{m,k} \quad B = \{b_{i,j}\}_{i,j=1}^{k,l} \quad C = \{c_{i,j}\}_{i,j=1}^{l,n}.
\]

Proof of associativity. The \((i, j)\)-th entry in \( D := (AB)C \) is

\[
d_{i,j} = \sum_{s=1}^{n} \left( \sum_{r=1}^{n} a_{i,r} \cdot b_{r,s} \right) \cdot c_{s,j}.
\]

Due to distributivity,

\[
d_{i,j} = \sum_{s=1}^{n} \left( \sum_{r=1}^{n} a_{i,r} \cdot b_{r,s} \cdot c_{s,j} \right) = \sum_{r=1}^{n} \left( a_{i,r} \cdot \sum_{s=1}^{n} b_{r,s} \cdot c_{s,j} \right),
\]

which is the \((i, j)\)-th entry in \( A(BC) \).

Therefore, \((AB)C = A(BC)\).

Proof of additivity. Again due to distributivity,

\[
\sum_{r=1}^{n} (a_{i,r} + b_{i,r}) \cdot c_{r,j} = \sum_{r=1}^{n} a_{i,r} \cdot c_{r,j} + \sum_{r=1}^{n} b_{i,r} \cdot c_{r,j}.
\]

Therefore, \((A + B)C = AC + BC\). The proof that \(A(B + C) = AB + AC\) is analogous.

Proof of homogeneity. Again due to distributivity,

\[
t \cdot \sum_{r=1}^{n} a_{i,r} \cdot b_{r,j} = \sum_{r=1}^{n} (t \cdot a_{i,r}) \cdot b_{r,j} = \sum_{r=1}^{n} a_{i,r} \cdot (t \cdot b_{r,j}).
\]

Therefore, \(t(AB) = (tA)B = A(tB)\). \(\square\)
Example 715. For $n > 1$, the matrix algebra $R^{n \times n}$ is a noncommutative ring. Consider the following example:

\[
\begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 \\
1 & 0
\end{pmatrix} =
\begin{pmatrix}
0 & 0 \\
1 & 0
\end{pmatrix},
\]

\[
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix} =
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix}.
\]

Remark 716. Let $R$ be a commutative ring and let $e_1, \ldots, e_n$ be the standard basis of $R$. The coordinate projections $\pi_{e_1}, \ldots, \pi_{e_n}$ allow us to identify $R^n$ with the module $R^{n \times 1}$ of column vectors by regarding the vector $x$ from $R^n$ as the column vector

\[
\begin{pmatrix}
\pi_{e_1}(x) \\
\vdots \\
\pi_{e_n}(x)
\end{pmatrix}.
\]

Under this identification, the columns on the identity matrix (230) are precisely the column vectors of the standard basis.

Let $A = \{a_{ij}\}_{i,j=1}^{m,n}$ be an $m \times n$ matrix over $R$. If we regard $R^n$ as a set of column vectors, then matrix multiplication allows us to regard $A$ as the function $x \mapsto Ax$, which maps column vectors from $R^n$ to column vectors in $R^m$.

This justifies using juxtaposition for application of linear maps, e.g. $Lx$ rather than $L(x)$. Conversely, let $e_1, \ldots, e_n$ be the standard basis of $R^n$ and $f_1, \ldots, f_m$ — of $R^m$. The linear map $L : R^n \to R^m$ corresponds to the following matrix:

\[
\begin{pmatrix}
\pi_{e_1}(L f_1) & \cdots & \pi_{e_1}(L f_m) \\
\vdots & \ddots & \vdots \\
\pi_{e_n}(L f_1) & \cdots & \pi_{e_n}(L f_m)
\end{pmatrix}.
\]

Proposition 717. For a commutative ring $R$, the matrix algebra $R^{m \times n}$ is isomorphic to the linear function algebra $\text{hom}(R^n, R^m)^7$.

Proof. Follows from our discussion in remark 716 due to linearity.

Remark 718. We want to be able to map single indices to double indices and vice versa, for example for the purpose of proposition 719. As an example, we want to be able to “linearize” an $m \times n$ matrix such as the $2 \times 3$ matrix

\[
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6
\end{pmatrix}
\]

into the tuple

\[
(1, 2, 3, 4, 5, 6)
\]

\[7\text{Note that the maps are from } R^n \text{ to } R^m \text{ and not vice versa.}\]
and vice versa. This is called **row-major order** of the elements of a matrix. The **column-major order** would instead be

\[(1, 4, 2, 5, 3, 6).\] (233)

Let \(m\) and \(n\) be positive integers. We will explicitly define functions for linearizing a matrix like (231) into its row-major order (232). Consider the sets

\[
S := \{1, \ldots, mn - 1, mn\} \\
D := \{1, \ldots, m\} \times \{1, \ldots, n\}
\]

and the mutually inverse operations

\[
\#: S \rightarrow D \\
\#(k) := (\text{quot}(k - 1, m) + 1, \text{rem}(k - 1, m) + 1)
\]

\[
\&: D \rightarrow S \\
\&(i, j) := (i - 1) \cdot m + (j - 1) + 1.
\]

The operation \# encodes the matrix (231) into its row-major order (232) and \& does the opposite. Both operations are trivial except for the shifting needed in to allow us to use remainders and quotients.

We can easily verify that \# is a left inverse of \& (note that \(j < m\)):

\[
\#(\&(i, j)) = \#((i - 1) \cdot m + (j - 1) + 1) = \\
= (\text{quot}(\cdots, m) + 1, \text{rem}(\cdots, m) + 1) = \\
= (i - 1) + 1, (j - 1) + 1 = \\
= (i, j).
\]

We can just as easily verify that \& is a right inverse of \#:

\[
\&(\#(k)) = \&(\text{quot}(k, m) + 1, \text{rem}(k, m) + 1) = \\
= \text{quot}(k, m) \cdot m + \text{rem}(k, m) = \\
= k.
\]

Hence, \# is fully invertible with inverse \&. By proposition 989 (f), it is bijective.

**Proposition 719.** The matrix algebra \(R^{mn}\) is isomorphic as a semimodule to \(R^{mn}\).

**Proof.** Remark 718 gives us a semimodule isomorphism between \(m \times n\) matrices and \(mn\)-dimensional column vectors when extended to linear maps via theorem 552 (Free semimodule universal property). \(\square\)
The determinant for the matrix algebra $R^{n \times n}$ over the commutative semiring $R$ is the function

$$\text{det} : R^{n \times n} \to R$$

$$\text{det}([a_{i,j}]_{i,j=1}^n) := \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n a_{i,\sigma(i)},$$

where $S_n$ is the symmetric group and $\text{sgn}$ is the sign of the permutation $\sigma$.

See the proof of proposition 725 for a justification of the definition.

Given a function $f : X^n \to Y$, where $X$ and $Y$ are plain sets, we say that $f$ is symmetric if, for any permutation $\sigma \in S_n$, we have

$$f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

A permutation can be decomposed into transpositions due to proposition 502. Hence, the above condition reduces to the simpler condition of $f$ being invariant with respect to swapping any two arguments. That is,

$$f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots) = f(\ldots, x_{i-1}, x_{i+1}, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots).$$

In the case where $n = 2$, this reduces to the simple condition

$$f(x, y) = f(y, x).$$

Symmetric functions should not be confused with symmetric binary relations defined in definition 954 (h).

Given a function $f : X^n \to Y$, where $X$ is a plain set and $Y$ is an additive group, we say that $f$ is antisymmetric if, for any permutation $\sigma \in S_n$, we have

$$f(x_1, \ldots, x_n) = \text{sgn}(\sigma) \cdot f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

A permutation can be decomposed into transpositions due to proposition 502. Hence, the above condition reduces to the simpler condition of $f$ changing sign when swapping any two arguments. That is,

$$f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots) = -f(\ldots, x_{i-1}, x_{i+1}, x_{i+1}, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots).$$

In the case where $n = 2$, this reduces to the simple condition

$$f(x, y) = -f(y, x).$$

Antisymmetric functions should not be confused with antisymmetric binary relations defined in definition 954 (i).

Given a commutative ring $R$, and $R$-module $M$ and a multilinear function $f : M \to R$, we say that $f$ is alternating if, $x_i = x_j$ implies that

$$f(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n) = 0.$$
Proposition 724. If a multilinear map is alternating, it is antisymmetric. The converse holds if 2 is a unit.

Proof.

Proof of sufficiency. If \( f \) is an alternating multilinear map, then

\[
0 = f(\cdots, x_i + x_j, \cdots, x_i + x_j, \cdots) = f(\cdots, x_i, \cdots, x_j, \cdots) + f(\cdots, x_j, \cdots, x_i, \cdots).
\]

Therefore,

\[
f(\cdots, x_i, \cdots, x_j, \cdots) = -f(\cdots, x_j, \cdots, x_i, \cdots).
\]

Proof of necessity. If \( f \) is an antisymmetric multilinear map, then

\[
0 = f(\cdots, x_i + x_i, x_i, \cdots) = 2f(\cdots, x_i, x_i, \cdots).
\]

If 2 is a unit, this implies

\[
f(\cdots, x_i, x_i, \cdots) = 0.
\]

Proposition 725. In the matrix algebra \( R^{n\times n} \) over the commutative ring \( R \), the determinant function \( \det : R^{n\times n} \to R \) can be regarded as a function that maps \( n \) column vectors from \( R^n \) to \( R \). That is,

\[
\det(v_1, \cdots, v_n) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^{n} \pi_{\sigma(i)}(v_i). \tag{237}
\]

The determinant is an alternating multilinear function on columns. Furthermore, it is the unique alternating multilinear function \( f(v_1, \cdots, v_n) \) such that \( f(e_1, \cdots, e_n) = 1 \), where \( e_1, \cdots, e_n \) are vectors of the standard basis in \( R^n \).

Proof. Let \( \pi_1, \cdots, \pi_n \) be the projection functionals corresponding to the standard basis \( e_1, \cdots, e_n \).

Proof of multilinearity. Due to linearity of the coordinate projection functionals \( \pi_i \) and due to distributivity in \( R \), for every \( j \) we have

\[
\det(\cdots, v_{j-1}, ty + rz, v_{j+1}, \cdots) =
\]

\[
\sum_{\sigma \in S_n} \text{sgn}(\sigma) \cdot \pi_{\sigma(j)}(ty + rz) \prod_{i \neq j} \pi_{\sigma(i)}(v_i) =
\]

\[
t \sum_{\sigma \in S_n} \text{sgn}(\sigma) \cdot \pi_{\sigma(j)}(y) \prod_{i \neq j} \pi_{\sigma(i)}(v_i) + r \sum_{\sigma \in S_n} \text{sgn}(\sigma) \cdot \pi_{\sigma(j)}(z) \prod_{i \neq j} \pi_{\sigma(i)}(v_i) =
\]

\[
t \cdot \det(\cdots, y, \cdots) + r \cdot \det(\cdots, z, \cdots).
\]

Proof of alternation. If \( v_i = v_j \), then for every even (resp. odd) permutation \( \sigma \), the permutation \((i\ j) \circ \sigma\) is odd (resp. even), and hence they cancel out in the sum (237). This holds for every permutation, hence it remains for the determinant to be zero.
Proof of \( \det(I_n) = 1 \). Note that

\[
\prod_{i=1}^{n} \pi_i(e_{\sigma(i)}) \neq 0
\]

if and only if \( i = \sigma(i) \) for every \( i = 1, \ldots, n \). This only holds for the identity permutation, hence

\[
\det(e_1, \ldots, e_n) = \prod_{i=1}^{n} \pi_i(e_i) = \prod_{i=1}^{n} 1 = 1.
\]

Proof of uniqueness. Suppose that \( f(v_1, \ldots, v_n) \) is an alternating multilinear function such that \( f(e_1, \ldots, e_n) = 1 \).

For an arbitrary column vector \( v_j \) in \( R^n \), we have

\[
v_j = \sum_{i=1}^{n} \pi_i(v_j) \cdot e_i.
\]

Then

\[
f(v_1, \ldots, v_n) = f \left( \sum_{i=1}^{n} \pi_i(v_1) \cdot e_i, \ldots, \sum_{i_n=1}^{n} \pi_{i_n}(v_n) \cdot e_{i_n} \right) =
\]

\[
= \sum_{i_1=1}^{n} \pi_i(v_1) \cdots \sum_{i_n=1}^{n} \pi_{i_n}(v_n) f(e_{i_1}, \ldots, e_{i_n}) =
\]

\[
= \sum_{\sigma \in S_n} \pi_{\sigma(1)}(v_1) f(e_{\sigma(1)}, \ldots, e_{\sigma(n)}).
\]

The last step is valid because \( f \) is alternating and thus \( f(e_{i_1}, \ldots, e_{i_n}) \) is zero when not all of \( i_1, \ldots, i_n \) are distinct, and they are necessarily distinct if the indices are given by a permutation from \( S_n \).

Finally, since \( f \) is antisymmetric due to proposition 724,

\[
f(e_{\sigma(1)}, \ldots, e_{\sigma(n)}) = \operatorname{sgn}(\sigma) f(e_1, \ldots, e_n) = \operatorname{sgn}(\sigma).
\]

Therefore,

\[
f(v_1, \ldots, v_n) = \det(v_1, \ldots, v_n).
\]

\[ \Box \]

Definition 726. The transpose matrix of

\[
A = \begin{pmatrix}
  a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
  a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{m,1} & a_{m,2} & \cdots & a_{m,n}
\end{pmatrix}
\]

355
is defined as

\[ A^T = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \]

A matrix that is equal to its transpose is called symmetric.

**Proposition 727.** In the matrix algebra \( R^{n \times n} \) over the commutative ring \( R \), the determinant as function on matrices has the following basic properties:

(a) \( \det(A^T) = \det(A) \).

(b) \( \det(tA) = t^n \cdot \det(A) \).

(c) \( \det(AB) = \det(A) \cdot \det(B) \).

That is, \( \det : R^{n \times n} \to R \) is a monoid homomorphism from the multiplicative monoid of the ring \( R^{n \times n} \) to the multiplicative monoid of \( R \).

**Proof.**

**Proof of 727 (a).** The inverse of any permutation in \( S_n \) is also a permutation in \( S_n \), hence

\[
\det(A^T) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i), j} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{i=1}^{n} a_{i, \sigma^{-1}(i)} = \det(A).
\]

**Proof of 727 (b).** Follows from proposition 725.

**Proof of 727 (c).** The \( j \)-th column of the product \( C = AB \) is

\[
c_{-j} = \sum_{i=1}^{n} b_{i,j} a_{-i} = \begin{pmatrix} \sum_{i=1}^{n} a_{1,i} b_{i,j} \\ \vdots \\ \sum_{i=1}^{n} a_{n,i} b_{i,j} \end{pmatrix}.
\]

Since the determinant is a multilinear function on columns,

\[
\det(c_{-1}, \cdots, c_{-n}) = \det\left( \sum_{i_1=1}^{n} b_{i_1,1} a_{-i_1}, \cdots, \sum_{i_n=1}^{n} b_{i_n,n} a_{-i_n} \right) = \\
= \sum_{i_1=1}^{n} b_{i_1,1} \det\left( a_{-i_1}, \cdots, \sum_{i_n=1}^{n} b_{i_n,n} a_{-i_n} \right) = \\
= \sum_{i_1=1}^{n} b_{i_1,1} \cdots \sum_{i_n=1}^{n} b_{i_n,n} \det(a_{-i_1}, \cdots, a_{-i_n}) = \\
= \sum_{i_1=1}^{n} \cdots \sum_{i_n=1}^{n} b_{i_1,1} \cdots b_{i_n,n} \det(a_{-i_1}, \cdots, a_{-i_n}).
\]
Since the determinant is alternating on columns, \( \text{det}(a_{-l_1}, \cdots, a_{-l_n}) \) is zero when not all of \( l_1, \ldots, l_n \) are distinct. They are necessarily distinct if the indices are given by a permutation from \( S_n \). Therefore,

\[
\begin{align*}
\text{det}(AB) &= \sum_{\sigma \in S_n} \prod_{i=1}^n a_{i,\sigma(i)} \cdot \sigma(a_{\sigma(1)}, \ldots, a_{\sigma(n)}) \\
&= \sum_{\sigma \in S_n} \prod_{i=1}^n a_{i,\sigma(i)} \cdot \text{sgn}(\sigma) \cdot \sigma(a_{-1}, \ldots, a_{-n}) \\
&= \text{det}(B) \text{det}(A).
\end{align*}
\]

\( \square \)

**Definition 728.** If for the matrices \( A = \{a_{i,j}\}_{i,j=1}^{m,n} \) and \( B = \{b_{i,j}\}_{i,j=1}^{k,k} \) over a commutative ring there exist monotone functions

\( h : \{1, \ldots, k\} \to \{1, \ldots, m\}, \)

\( w : \{1, \ldots, l\} \to \{1, \ldots, n\}, \)

such that, for every \( i = 1, \ldots, k \) and \( j = 1, \ldots, l \) we have

\( b_{i,j} = a_{h(i),w(j)}. \)

**Definition 729.** A **minor** of a matrix is a determinant of a square submatrix.

**Theorem 730** (Laplace expansion). For a square matrix \( A = \{a_{i,j}\}_{i,j=1}^{n,n} \) over a commutative ring and a row index \( i \), we have

\[
\text{det} A = \sum_{j=1}^n (-1)^{i+j} a_{i,j} \text{det} A_{i,j},
\]

where \( A_{i,j} \) is the submatrix of \( A \) obtained by removing the \( i \)-th row and the \( j \)-th column.

By proposition 727 (a), we can also expand along a column rather than a row.

**Proof.** Denote the ring by \( R \). We will show that, for the \( i \)-th row,

\[
\Phi : R^{n \times n} \to R,
\]

\[
\Phi(A) := \sum_{j=1}^n (-1)^{i+j} a_{i,j} \text{det} A_{i,j},
\]

is an alternating multilinear function on columns.

Multilinearity follows from the multilinearity of determinants. For proving alternation, suppose that the \( k \)-th and \( l \)-th columns are equal. Then

\[
\Phi(A) = (-1)^{i+k} a_{i,k} \text{det} A_{i,k} + (-1)^{i+l} a_{i,l} \text{det} A_{i,l}.
\]
The matrix $A_{i,l}$ can be obtained from $A_{i,k}$ by swapping $|k-l|$ columns. Since determinants are antisymmetric, it follows that

$$\det A_{i,l} = (-1)^{k-l} \det A_{i,k}.$$ 

Furthermore, $a_{i,k} = a_{i,l}$. Therefore,

$$\Phi(A) = (-1)^{i+k} a_{i,k} \det A_{i,k} + (-1)^{(i+l)+(k-l)} a_{i,k} \det A_{i,k} = 0. \qed$$

**Definition 731.** The cofactor matrix of the $m \times n$ matrix $A$ is

$$\{(−1)^{i+j} \det A_{i,j}\}_{i,j=1}^{m,n},$$

where $A_{i,j}$ is the submatrix of $A$ obtained by removing the $i$-th row and the $j$-th column.

The adjugate matrix, also called the classical adjoint matrix, is the transpose of the cofactor matrix.

**Proposition 732.** The adjugate matrix of the square $n \times n$ matrix $A$ satisfies

$$A \cdot A^{\text{adj}} = \det A \cdot I_n.$$ 

**Proof.** From theorem 730 (Laplace expansion) it follows that the $(i,i)$-th entry of the matrix $A \cdot A^{\text{adj}}$ is

$$\sum_{k=1}^{n} (-1)^{i+k} a_{i,k} \det A_{i,k} = \det A.$$ 

For $i \neq j$, the $(i,j)$-th entry is

$$\sum_{k=1}^{n} (-1)^{i+j} a_{i,k} \det A_{j,k} = \det A_{j-i},$$

where $A_{j-i}$ is the matrix obtained by replacing the $j$-th column in $A$ with the $i$-th. The determinant is then zero because it is an alternating function on the columns.

Thus, the proposition follows. $\square$

**Definition 733.** We say that $B \in R^{n \times m}$ is a left inverse matrix of $A \in R^{m \times n}$ if $BA$ is the identity matrix $I_n$ and a right inverse matrix if $AB$ is $I_m$. These are precisely the left and right inverse linear maps in the correspondence described in proposition 717.

Due to proposition 734, for square matrices, the two notions coincide, and we say that $B$ is simply an inverse of $A$. An inverse matrix, if it exists, is unique. We denote this inverse of $A$ by $A^{-1}$.

We say that $A$ is invertible if an inverse exists, and singular otherwise.

**Proof of correctness.** The inverse is unique by proposition 453. $\square$
Proposition 734. Over a nontrivial noetherian commutative ring $R$, a square matrix is left invertible if and only if it is right invertible.

Proof.

Proof of necessity. Suppose that $A$ is a right invertible matrix. When regarding $A$ as a linear map via the identification from remark 716, this implies that $A$, as a linear map from $R^n$ to $R^n$, is right invertible. Then it is surjective and, by proposition 658, an isomorphism. Therefore, $A$ is a fully invertible matrix.

Proof of sufficiency. Now suppose that $A$ is a left inverse of $B$. Then $B$ is a right inverse of $A$, and, by the other direction of the proposition, a two-sided inverse of $A$. $\square$

Proposition 735. In the matrix algebra $R^{n\times n}$ over a nontrivial noetherian commutative ring $R$, the following are equivalent:

(a) The matrix $A$ is invertible.

(b) The determinant of $A$ is invertible.

(c) The columns of $A$ are linearly independent.

Proof.

Proof that 735 (a) implies 735 (b). Suppose that $A$ is invertible.

By proposition 727 (c),

$$\det(A^{-1})\det(A) = \det(A^{-1}A) = \det(I_n) = 1,$$

hence $\det(A)$ has a multiplicative inverse.

Proof that 735 (b) implies 735 (c). As in proposition 725, regard $\det(v_1, \ldots, v_n)$ are an alternating multilinear function on the columns of a matrix.

Suppose that $\det(v_1, \ldots, v_n)$ is a unit in $R$ and, aiming at a contradiction, suppose that the column vectors $v_1, \ldots, v_n$ are linearly dependent. Then there exists a nontrivial linear combination that sums to zero:

$$\sum_{i=1}^n t_i v_i = 0.$$

Suppose that $t_k$ is nonzero. Then

$$0 = \det(v_1, \ldots, v_{k-1}, 0, v_{k+1}, \ldots, v_n) =$$

$$= \det(v_1, \ldots, v_{k-1}, \sum_{i=1}^n t_i v_i, v_{k+1}, \ldots, v_n) =$$

$$= \sum_{i=1}^n t_i \det(v_1, \ldots, v_{k-1}, v_i, v_{k+1}, \ldots, v_n) =$$

$$= t_k \det(v_1, \ldots, v_{k-1}, v_k, v_{k+1}, \ldots, v_n).$$

But we have assumed that $\det(v_1, \ldots, v_{k-1}, v_k, v_{k+1}, \ldots, v_n)$ is a unit and that $t_k$ is nonzero. Hence, the determinant can only be a zero divisor if the ring is trivial, which we have assumed it is not. The obtained contradiction shows that $v_1, \ldots, v_n$ are linearly independent.
Proof that 735 (c) implies 735 (a). Suppose that the columns of $A$ are linearly independent. Consider the matrix equation

$$Ax = \begin{pmatrix} a_{-1} & \cdots & a_{-n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{k=1}^{n} x_k a_{-k} = \vec{0}.$$  

Since the columns are linearly independent, only $x_1 = \cdots = x_n$ is a solution to this equation. Thus, when regarding $A$ as the linear map $x \mapsto Ax$, the kernel of $A$ becomes trivial. By proposition 457 (g), this map is injective. As discussed in definition 604 (f), the injective linear maps are exactly the left invertible linear maps. Hence, there exists a left inverse of $A$. Since $A$ is a square matrix, by proposition 734, this implies that $A$ is invertible. \hfill \Box
11.2. Matrices over fields

We will assume that all matrices have entries from some fixed field $\mathbb{K}$. We will later on need to distinguish between real and complex matrices, but the theory built here holds more generally than that, and we choose to postulate it for arbitrary fields.

The definitions of triangular and elementary matrices make sense over more general rings, however we introduce them because of algorithm 740 (PLU decomposition), which has no direct generalization.

**Definition 736.** An upper triangular matrix is one with zeros below its main diagonal. More precisely, $U = \{u_{i,j}\}_{i,j=1}^{m,n}$ is an upper triangular matrix if $u_{i,j} = 0$ when $i > j$.

Similarly, a lower triangular matrix is one with zeros above its main diagonal.

A matrix that is either upper or lower triangular is simply referred to as “triangular”.

**Proposition 737.** Triangular matrices have the following basic properties:

(a) A matrix that is both upper and lower triangular is a diagonal matrix.

(b) The product of upper (resp. lower) triangular matrices is upper (resp. lower).

Consequently, the product of diagonal matrices is a diagonal matrix.

(c) The determinant of a triangular matrix is the product of (the entries on) its main diagonal.

(d) A triangular matrix is invertible if and only if its main diagonal has no zero entries.

Here, the assumption that $\mathbb{K}$ is a field is essential.

**Proof.**

**Proof of 737 (a).** Trivial.

**Proof of 737 (b).** Let $A$ be an $m \times k$ upper triangular matrix and $B$ be a $k \times n$ upper triangular matrix. The $(i,j)$-th element of $C = AB$ is

$$
\sum_{l=1}^{k} a_{i,l} b_{l,j}.
$$

Since $A$ and $B$ are upper triangular, we have $b_{l,j} = 0$ whenever $l > j$ and $a_{i,l} = 0$ whenever $l < i$. Thus, $a_{i,j} b_{l,j} = 0$ if either condition holds. If $l > j$, then either $l > j$ or $l < j < l$, implying that $a_{i,l} b_{l,j} = 0$. Therefore, $AB$ is also upper triangular.

The proof for lower triangular matrices is analogous.

**Proof of 737 (c).** Let $A$ be an $n \times n$ upper triangular matrix. Let $\sigma \in S_n$ be any permutation. Then $a_{i,\sigma(i)} = 0$ when $i > \sigma(i)$. Hence, the only permutation for which the product $\prod_{i=1}^{n} a_{i,\sigma(i)}$ is nonzero is the identity permutation. Therefore,

$$
\det(A) = \prod_{i=1}^{n} a_{i,i}.
$$
Proof of 737 (d). Follows from proposition 737 (c) and proposition 735 by noting that 0 is the only non-invertible element in a field.

Definition 738. We introduce the following three types of invertible $n \times n$ matrices, collectively known as elementary matrices:

(a) For a permutation $\sigma \in S_n$, the permutation matrix $P_\sigma$ is obtained by permuting the columns $e_1, \ldots, e_n$ of the identity matrix $I_n$ in accordance with $\sigma$. The permutation matrix

$$P_\sigma = \begin{pmatrix} e_{\sigma(1)} & \cdots & e_{\sigma(n)} \end{pmatrix}$$

acts on the $n \times m$ matrix $B = \{b_{i,j}\}_{i,j=1}^{m,n}$ by permuting the rows of $B$, i.e.

$$P_\sigma B = \begin{pmatrix} b_{\sigma(1),1} & \cdots & b_{\sigma(1),m} \\ \vdots & \ddots & \vdots \\ b_{\sigma(n),1} & \cdots & b_{\sigma(n),m} \end{pmatrix} \cdot \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m} \\ \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,m} \end{pmatrix} = \begin{pmatrix} b_{1,\sigma(1)} & b_{1,\sigma(1),2} & \cdots & b_{1,\sigma(1),m} \\ \vdots & \ddots & \vdots \\ b_{n,\sigma(n),1} & b_{n,\sigma(n),2} & \cdots & b_{n,\sigma(n),m} \end{pmatrix}.$$

The inverse is the matrix corresponding to its inverse permutation.

(b) For a nonzero element $a$ and index $i$, the scaling matrix $S_{i \mapsto a}$ is a diagonal matrix that differs from the identity by replacing 1 with $a$ instead of 1 in the $(i, i)$-th place. The scaling matrix

$$S_{i \mapsto a} := \begin{pmatrix} 1 & \cdots & i & i+1 & \cdots & n \\ 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \ddots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & a & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

acts on the $n \times m$ matrix $B$ by scaling the $i$-th row of $B$ by $a$.

The inverse is the same matrix with $a$ replaced by its multiplicative inverse $a^{-1}$.

(c) For any element $a$ and indices $i$ and $j$, the transvection matrix $T_{i \mapsto a,j}$ is obtained from the identity matrix $I_n$ by placing $a$ on the $(j, i)$-th place. The transvection matrix

$$T_{i \mapsto a,j} := \begin{pmatrix} 1 & \cdots & i & n \\ 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & a & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \cdots & 0 & 1 \\ j & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

362
acts on the \( n \times m \) matrix \( B \) by adding the \( i \)-th row of \( B \) scaled by \( a \) to the \( j \)-th row.

The order of indices is important — if the \((j, i)\)-th entry is nonzero, the scaled \( i \)-th row gets added to the \( j \)-th.

The inverse is the same matrix with \( a \) replaced by its additive inverse \(-a\).

**Proposition 739.** Elementary matrices have the following basic properties:

(a) The product of permutation matrices is a permutation matrix.

(b) The determinant of a permutation matrix is the sign of the permutation.

(c) The product of the transvection matrices \( T_{i, \alpha} \) and \( T_{i, \beta} \) is the transvection matrix \( T_{i, \alpha + \beta} \).

(d) The product of the transvection matrices \( A = T_{i_A, \alpha} \) and \( B = T_{i_B, \beta} \) with \( i_A \neq i_B \) or \( j_A \neq j_B \) is the identity matrix \( I_n \) modified with \( \alpha \) in the \((j_A, i_A)\)-th place and \( \beta \) in the \((j_B, i_B)\)-th.

**Proof.**

**Proof of 739 (a).** Trivial.

**Proof of 739 (b).** As discussed in the proof of proposition 725, the determinant of any permutation of the vectors of the standard basis is the sign of the permutation.

**Proof of 739 (d).** The \((i, j)\)-th entry of the product \( C = AB \) is

\[
c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}.
\]

- If \( i = j \), \( c_{i,j} \) is clearly 1.
- If \( i = i_A \) and \( j = j_A \), then

\[
a_{i_A,k} b_{k,j_A} = \begin{cases} a_{i_A,j_A}, & i_A = j_A, \\ 0, & i_A \neq j_A. \end{cases}
\]

Thus, \( a_{i_A,j_A} = a_{i_A,j_A} \).

- Analogously, \( c_{i_A,j_B} = b_{i_A,j_B} \).

- Otherwise, for \( k = 1, \ldots, n \), either \( a_{i,k} \) or \( b_{k,j} \) is zero, hence \( c_{i,j} \) also is.

**Proof of 739 (c).** This proof only requires a slight modification to the proof of proposition 739 (d).

**Algorithm 740 (PLU decomposition).** Fix an \( n \times n \) matrix \( A \). We will build a lower triangular matrix \( L \), upper triangular matrix \( U \) and a permutation matrix \( P \) such that \( A = PLU \).

The algorithm itself is also called Gaussian elimination and \( U \) is said to be a row-echelon form of \( A \), although both terms may have different meanings depending on the context.

363
We will proceed via bounded recursion on \( n \). After the \( k \)-th step, for \( k = 1, \ldots, n - 1 \), we will have built a lower triangular matrix \( L_k \) and a permutation matrix \( P_k \) such that for \( i > k \), \((i, k)\)-th entry of \( L_k P_k A \) is zero.

Furthermore, we will obtain \( L_k \) as a product of transvection and permutation matrices. Therefore, at each step, both \( P_k \) and \( L_k \) will be invertible as products of invertible matrices.

The matrix \( U := L_{n-1} P_{n-1} A \) will be upper triangular, and hence, putting \( P := P_{n-1} \) and \( L := L_{n-1}^{-1} \), we obtain
\[
A = P L U.
\]

(a) As an initial condition, put \( L_0 := I_n \) and \( P_0 := I_n \) as identity matrices.

(b) Suppose that we have already built \( L_{k-1} \) and \( P_{k-1} \). Let \( U_{k-1} := L_{k-1} P_{k-1} A \). We will describe step \( k \) of the algorithm.

If the \((k, j)\)-th entry of \( U_{k-1} \) is zero for all \( j > k \), put \( P_k = P_{k-1} \) and \( L_k = L_{k-1} \).

Otherwise, let \( j_0 \) be the first row index of \( L_{k-1} P_{k-1} A \) for which the \( k \)-th entry is nonzero. Let \( P_{k \rightarrow j_0} \) be the permutation matrix exchanging the \( k \)-th and \( j_0 \)-th column of the identity and put
\[
P_k := P_{k \rightarrow j_0} P_{k-1}.
\]

Then, since \( P_{k \rightarrow j_0} \) is its own inverse,
\[
\tilde{U}_{k-1} := P_{k \rightarrow j_0} U_{k-1} = P_{k \rightarrow j_0} L_{k-1} (P_{k \rightarrow j_0} P_{k-1} A) = (P_{k \rightarrow j_0} L_{k-1} P_{k \rightarrow j_0}) P_k A.
\]

Denote by \( u_{i,j} \) the entries of \( \tilde{U}_{k-1} \).

Also put \( \hat{L}_{k-1} := P_{k \rightarrow j_0} L_{k-1} P_{k \rightarrow j_0} \). This is again a lower triangular matrix since we only swap columns below the main diagonal.

For each row \( j > k \), define \( v_j := -u_{k,j}/u_{k,k} \) consider the transvection matrix \( T_{k \rightarrow j} \).

When multiplied by \( \tilde{U}_{k-1} \) from the right, it adds the \( k \)-th row of \( \tilde{U}_{k-1} \) to the \( j \)-th after multiplying it by \( v_j \). Hence, \( T_{k \rightarrow j} \tilde{U}_{k-1} \) has zero at as its \((i, j)\)-th entry.

Finally, put
\[
L_k := \left( \prod_{j=k}^{n} T_{k \rightarrow j} \right) \hat{L}_{k-1}.
\]

By proposition 739 (d), \( L_k \) adds nonzero entries to \( \hat{L}_{k-1} \) only below the main diagonal. Since \( \hat{L}_{k-1} \) is lower triangular, so is \( L_k \). Furthermore, for \( j > k \), the coefficient \( v_j \) is chosen so that the \((k, j)\)-th entry of \( L_k P_k A \) of zero, which ensures that the latter matrix will be upper triangular when \( k = n - 1 \).

**Proposition 741.** Let \( A = P L U \) be the decomposition of some matrix \( A \) obtained via algorithm 740 (PLU decomposition).

(a) If \( A \) is upper triangular, then \( P = L = I_n \) and \( A = U \).
(b) $A$ is nonsingular if and only if $U$ is nonsingular.

**Proof.**

**Proof of 741 (a).** Suppose that $A$ is upper triangular. Then, at step $k$ of the algorithm:

- If $u_{kk}$ is zero, then all entries below it are also zero, and hence we directly continue to the next step.
- If $u_{kk}$ is not zero, then the transvection matrices $T_{k \rightarrow j}$ for $j > k$ are all identity matrices, and hence $L_k$ is also the identity.

In both cases, $L_k = P_k = I_n$.

**Proof of 741 (b).** By proposition 727 (c),

$$\det(A) = \det(P) \det(L) \det(U).$$

Since $P$ and $L$ are products of permutation and transvection matrices, by ?? and proposition 739 (d), their determinants are either 1 or $-1$. Hence,

$$|\det(A)| = |\det(U)|.$$

It follows from proposition 739 (d) that $A$ is nonsingular if and only if $U$ is. \qed

**Algorithm 742 (Elementary matrix decomposition).** Fix a nonsingular $n \times n$ matrix $A$. Let $A = PLU$ be the decomposition obtained via algorithm 740 (PLU decomposition). By proposition 741 (b), $U$ is a nonsingular matrix. Both $P$ and $L$ are products of elementary matrices, hence it suffices to show that $U$ is a product of elementary matrices in order to show that $A$ is a product of elementary matrices.

The algorithm is complementary to algorithm 740 (PLU decomposition), although with noticeable differences. We will assume that $k = 2, \ldots, n$. At each step, we will build a matrix $U_k$ whose first $k$ columns match those of $U$. Then $U_n$ must equal $U$.

Denote by $u_{ij}$ the entries of $U$.

(a) We will define the initial condition $U_1$ to be the diagonal matrix whose main diagonal matches that of $U$. This can be achieved via scaling matrices:

$$U_1 := \prod_{i=1}^{n} S_{i \rightarrow u_{ii}}.$$

(b) At step $k$, given $U_{k-1}$, for $j < k$ define $v_j := (u_{k,j})/(u_{k,k})$. It is important that here, unlike in algorithm 740 (PLU decomposition), we put no minus sign in $v_j$ since we are building the matrix $U$ directly rather than building an intermediate matrix that we will later invert. Put

$$U_k := \left( \prod_{j=k}^{n} T_{k \rightarrow j} v_j \right) U_{k-1}.$$
Since $U$ is nonsingular, by proposition 737 (c), $u_{k,k}$ must be nonzero. Hence, we can divide by it.

As a product of scaling and transvection matrices with nonzero entries above the main diagonal, $U_k$ is an upper diagonal matrix. Furthermore, the scaling matrices neutralize the division done by the transvection matrices, so the $k$-th column of $U_k$ and $U$ must match.

**Proposition 743.** The square matrix over a field is invertible if and only if it is a product of elementary matrices.

**Proof.**

**Proof of sufficiency.** Follows from algorithm 742 (Elementary matrix decomposition).

**Proof of necessity.** Follows from proposition 727 (c).

**Example 744.** Given elements $r_0, r_1, \ldots, r_n$ of some commutative ring, we define their Vandermonde matrix as

$$V_n(r_0, r_1, \ldots, r_n) := \begin{pmatrix}
    r_0 & r_1^1 & r_1^2 & \cdots & r_1^n \\
    r_0^1 & r_0 & r_0^2 & \cdots & r_0^n \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    r_n^0 & r_n^1 & r_n^2 & \cdots & r_n^n 
\end{pmatrix}.$$

Having in mind that, by proposition 737 (c), transvection matrices have determinant 1, subtracting the $k$-th row multiplied by $r_0$ from the $(k+1)$-th does not change the determinant. We thus have

$$\det V_n = \det V_n^T = \det \begin{pmatrix}
    1 & 1 & 1 & \cdots & 1 \\
    0 & r_1 - r_0 & r_2 - r_0 & \cdots & r_n - r_0 \\
    0 & r_1(r_1 - r_0) & r_2(r_2 - r_0) & \cdots & r_n(r_n - r_0) \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    0 & r_n^{k-1}(r_1 - r_0) & r_n^{k-1}(r_2 - r_0) & \cdots & r_n^{k-1}(r_n - r_0) \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    0 & r_n^{n-1}(r_1 - r_0) & r_n^{n-1}(r_2 - r_0) & \cdots & r_n^{n-1}(r_n - r_0) 
\end{pmatrix}$$

$$= \det \begin{pmatrix}
    r_1 - r_0 & r_2 - r_0 & \cdots & r_n - r_0 \\
    r_1(r_1 - r_0) & r_2(r_2 - r_0) & \cdots & r_n(r_n - r_0) \\
    \vdots & \vdots & \ddots & \vdots \\
    r_1^{k-1}(r_1 - r_0) & r_2^{k-1}(r_2 - r_0) & \cdots & r_n^{k-1}(r_n - r_0) \\
    \vdots & \vdots & \ddots & \vdots \\
    r_1^{n-1}(r_1 - r_0) & r_2^{n-1}(r_2 - r_0) & \cdots & r_n^{n-1}(r_n - r_0) 
\end{pmatrix}$$
\[(r_1 - r_0)(r_2 - r_0) \cdots (r_n - r_0) \det \begin{pmatrix}
1 & 1 & \cdots & 1 \\
r_1 & r_2 & \cdots & r_n \\
\vdots & \vdots & \ddots & \vdots \\
r_{k-1} & r_{k-1} & \cdots & r_{k-1} \\
r_1 & r_2 & \cdots & r_{n-1}
\end{pmatrix}.
\]

Proceeding by induction, we conclude that
\[
\det V_n = \prod_{i<j} (r_j - r_i).
\]

Hence, the determinant is nonzero if and only if all of \(r_0, \ldots, r_n\) are distinct.

**Definition 745.** The **column space** of the \(m \times n\) matrix \(A\) is the linear span of the columns vectors of \(A\), regarded as a subspace of \(\mathbb{K}^n\). It is precisely the image of \(A\) regarded as a linear operator.

Analogously, the **row space** is the span of the row vectors, regarded as a subspace of \(\mathbb{K}^m\). It is the image of \(A^T\) regarded as a linear operator.

**Proposition 746.** The column space of a matrix is isomorphic to its row space.

We call the dimension of these spaces the **rank** of the matrix. We say that the matrix has **full rank** if its rank matches the smaller of its dimensions.

**Proof.** Let \(A\) be an \(m \times n\) matrix over \(\mathbb{K}\) and let \(r\) be the dimension of the image of \(A\). By **Theorem 606** (Quotient module universal property), as a linear operator, \(A\) factors through
\[
\mathbb{K}^m / \ker A \cong \text{img} A \cong \mathbb{K}^r.
\]

More precisely, there exists an \(m \times r\) matrix \(B\) and an \(r \times n\) matrix \(C\) such that \(A = BC\).

If \(A = BC\), then \(A^T = C^TB^T\). We cannot, at this point, rule out that \(r\) is larger than the dimension of \(\text{img} A^T\). Hence,
\[
\dim \text{img} A^T \leq \dim \text{img} A.
\]

Conversely,
\[
\dim \text{img} A = \dim \text{img}(A^T)^T \leq \dim \text{img} A^T.
\]

Therefore, the dimensions of the images of \(A\) and \(A^T\) are equal, and hence the images are isomorphic. \(\square\)

**Remark 747.** Given any map \(T : \mathbb{K}^n \to \mathbb{K}^m\) and an \(m\)-dimensional column vector \(b\), consider the **system of equations**
\[
T(x) = b.
\]

A **solution** to this system is an \(n\)-dimensional **column vector** \(x\). When regarding \(x\) as a free variable in the sense of **Definition 830**, the entries of \(x\) are called the **variables** of the system. When regarding \(T\) has a column vector of functions (via the **projection functionals**),
each individual row is called a **equation** of the system. The above is thus a system with $n$ variables and $m$ equations.

The system is called **homogeneous** if $b$ is the zero vector. To every non-homogeneous system, there corresponds a homogeneous system

$$ T(x) = 0. $$ (240)

In the case where $T$ is a linear function, we say that the system itself is linear. We can, in that case, rewrite the system via the matrix of $T$ with respect to the canonical bases. A linear system is often written in scalar form as

$$
\begin{align*}
& a_{1,1}x_1 + \cdots + a_{1,n}x_n = b_1 \\
& \vdots \\
& a_{m,1}x_1 + \cdots + a_{m,n}x_n = b_m.
\end{align*}
$$ (241)

If $x$ is a solution to a general linear system $Ax = b$ and if $y$ is a solution to the homogeneous system $Ax = 0$, then $x + y$ is also a solution to the general system. The proof is trivial:

$$ A(x + y) = Ax + Ay = Ax. $$

The theory of systems of linear equations precedes matrix theory. In applications, this is often a more convenient framework than the one provided (only) by matrices.

**Theorem 748** (Kroneker-Capelli theorem). The system

$$ Ax = b $$

of $n$ variables and $m$ linear equations has a solution if and only if the rank of its coefficient matrix $A$ is equal to the rank of its augmented matrix

$$ ( A | b ). $$

**Proof.** If both matrices have the same rank, $b$ belongs to the column space of $A$. Then there exists some linear combination of the columns of $A$ the equal $b$. The coefficients of every such linear combination form a solution to the system. \[\square\]
11.3. Bilinear forms

**Definition 749.** Let $M$ and $N$ be left $R$-modules and $L : M \times N \to R$ be a multilinear function. We say that $L$ is a **bilinear form**.

If $M = N$, we have the following additional types of bilinear forms:

(a) If $L$ is a symmetric function, we say that $L$ is a **symmetric bilinear form**

(b) If for all $x, y \in M$ instead of $L(x, y) = L(y, x)$ we have $L(x, y) = -L(y, x)$, we say that $L$ is **skew-symmetric**.

(c) If for all $x \in M$ we have $L(x, x) = 0$, we say that $L$ is **alternating**.

**Proposition 750.** Let $1 + 1 = 2$ be a unit in the ring $R$. Let $M$ be a left module over $R$. Then the bilinear form $L : M \times M \to R$ is alternating if and only if it is skew-symmetric. Alternating implies skew-symmetric even if 2 is not invertible.

**Proof.**

**Proof of sufficiency.** Let $L$ be alternating. Then

\[
L(x + y, x + y) = L(x, x) + L(x, y) + L(y, x) + L(y, y) = 0 = L(x, y) + L(y, x)
\]

so $L(x, y) = -L(y, x)$.

**Proof of necessity.** Let $L$ be skew-symmetric. Then

\[
L(x, x) = -L(x, x),
\]

which implies that $2L(x, y) = 0$. Hence, $L$ is alternating if we are able to divide by 2. \qed

**Definition 751.** Let $L : M \times N \to R$ be a bilinear form. We define its **left radical**

\[
\{x \in M : \forall y \in N, \langle x, y \rangle = 0\}
\]

and **right radical**

\[
\{y \in N : \forall x \in M, \langle x, y \rangle = 0\}.
\]

Note that if $L$ is symmetric or skew-symmetric (which also implies $M = N$), the two are identical and we speak simply of the **radical** $\sqrt{L}$.

**Definition 752.** We say that a bilinear form $L : M \times N \to R$ is **nondegenerate** if both its left and right radicals are nontrivial.

**Theorem 753.** Fix a commutative unital ring $R$ and a bilinear form $L : R^n \times R^m \to R$. Then there exists a matrix $A \in R^{nxm}$ such that

\[
L(x, y) := x^T Ay.
\]

This matrix is called the **generalized Gram matrix**. In particular, if $L$ is symmetric, so it $A$. 369
Proof. Denote by $e_1, \ldots, e_n$ the basis of $R^n$ and by $f_1, \ldots, f_m$ the basis of $R^m$. Define the matrix $A = \{a_{i,j}\}_{i,j=1}^{n,m}$ by

$$a_{i,j} := L(e_i, f_j).$$

Note that if $n = m$ and if $L$ is symmetric, then the matrix $A$ is obviously symmetric too. For any fixed basis vector $e_i, i = 1, \ldots, n$ of $R^n$, we have

$$L(e_i, y) = \sum_{j=1}^m y_j L(e_i, f_j) = y_i a_{i(1,-)},$$

where $a_{i(1,-)}$ is the $i$-th row of $A$. Thus, for an arbitrary $x \in R^n$

$$L(x, y) = \sum_{i=1}^n x_i L(e_i, y) = \sum_{i=1}^n x_i (a_{i(1,-)} y) = \left( \sum_{i=1}^n x_i a_{i(1,-)} \right) y = x^T A y.
\qed$$

**Corollary 754.** Fix a commutative unital ring $R$. The vector space of bilinear forms of type $R^n \times R^m \to R$ is isomorphic to the matrix space $A \in R^{n \times m}$.

**Definition 755.** Let $V$ be a complex vector space and let $\overline{V}$ be its conjugate transpose. We call the bilinear form $L : V \times \overline{V} \to \mathbb{C}$ a **sesquilinear form** (we say that $L$ is “semilinear” in its second argument and “sesqui” means “one and a half” in Latin).

Similar to **definition 749**, we have

(a) If for all $x, y \in V$ we have $L(x, y) = \overline{L(y, x)}$, we say that $L$ is **Hermitian**.

(b) If for all $x, y \in V$ we have $L(x, y) = -\overline{L(y, x)}$, we say that $L$ is **skew-Hermitian**.

**Definition 756.** If $L : M \times M \to R$ be a bilinear form, we call the function

$$Q(x) := L(x, x)$$

a **quadratic form** over $M$.

**Definition 757.** **TODO**: Define definiteness of quadratic forms

**Definition 758.** Let $M$ and $N$ be left $R$-modules. We say that the function $f : M \to N$ is homogeneous with degree $n$ if for all $t \in R$ and $x \in M$ we have

$$f(tx) = t^n f(x).$$

**Proposition 759.** A **quadratic form** $Q : M \to R$ is a homogeneous function of degree 2. In particular, $Q(x) = Q(-x)$. 370
Proof. Let $L : M \times M \to R$ be the corresponding bilinear form. Then, by eq. (198),

$$Q(tx) = L(tx, tx) = t^2L(x, x) = t^2Q(x).$$

□

Proposition 760. Let $L : M \times M \to R$ be a bilinear form and $Q : M \to R$ be its associated quadratic form. Then the polarization identity holds:

$$2L(x, y) + 2L(y, x) = Q(x + y) - Q(x - y)$$

The similar looking, but slightly less useful parallelogram law also holds:

$$2Q(x) + 2Q(y) = Q(x + y) + Q(x - y)$$

If $2 = 1 + 1$ is a unit in $R$, we can “recover” from $Q$ the bilinear form:

$$\hat{L}(x, y) := \frac{1}{2} [Q(x + y) - Q(x) - Q(y)]$$

The function $\hat{L}$ is symmetric and is called the symmetrization of $L$. If $L$ itself is symmetric, $L = \hat{L}$.

Proof. Identities eqs. (242) to (244) all follow from the bilinearity of $L$, that is,

$$Q(x \pm y) = L(x, x) \pm L(x, y) \pm L(y, x) + L(y, y) = [Q(x) + Q(y)] \pm [L(x, y) + L(y, x)].$$

□

Definition 761. Let $U$ and $V$ be vector spaces over $\mathbb{K}$ and let $L : U \times V \to \mathbb{K}$ be a non-degenerate bilinear form. We say that the vectors $x \in U$ and $y \in V$ are orthogonal with respect to $L$ if

$$L(x, y) = 0.$$ 

For every subspace $U \subseteq V$ we define its orthogonal complement with respect to $L$ as

$$U^\perp := \{x \in U : L(x, y) = 0 \text{ for all } y \in V\}$$

and analogously for submodules of $V$.

Let $\mathcal{X}$ be an index set and $\{x_k\}_{k \in \mathcal{X}} \subseteq U$, $\{y_k\}_{k \in \mathcal{X}} \subseteq V$ be two families of vectors indexed by $\mathcal{X}$. We say that these families form a biorthogonal system with respect to $L$ if

$$L(x_k, y_m) = 0 \text{ follows from } k \neq m$$

If $U = V$, we usually consider orthogonal systems $\{x_k\}_{k \in \mathcal{X}} \subseteq V$ where

$$L(x_k, x_m) = 0 \iff k \neq m$$

371
Definition 762. An inner product space is a vector space \( V \) over \( F \) equipped with a positive definite symmetric bilinear form \( \langle \cdot, \cdot \rangle : V \times V \to F \).

In the special case where \( F = \mathbb{C} \), by convention we require \( V \) to instead be equipped with a positive definite Hermitian sesquilinear form instead.

Definition 763. A symplectic vector space is a vector space \( V \) over \( F \) equipped with a nondegenerate alternating bilinear form \( \langle \cdot, \cdot \rangle : V \times V \to F \).

Lemma 764. Let \( V \) be a real or complex inner product space with product \( \langle \cdot, \cdot \rangle \). The function \( Q(x) := \langle x, x \rangle \) (which is not a quadratic form in the complex case) is positive definite.

Proof. The real case is trivial. Assume that \( V \) is a complex vector space and that \( \langle \cdot, \cdot \rangle \) is Hermitian. This implies that \( \langle x, x \rangle = \langle x, \bar{x} \rangle \), thus \( \langle x, x \rangle \in \mathbb{R} \). Furthermore, since the inner product is positive definite, we have \( Q(x) = \langle x, x \rangle \geq 0 \). Thus, \( Q \) is nonnegative real valued.

Since \( \langle \cdot, \cdot \rangle \) is positive definite, so is \( Q \). \( \square \)

Theorem 765 (Cauchy-Bunyakovsky-Schwarz inequality). Let \( V \) be a real or complex inner product space with product \( \langle \cdot, \cdot \rangle \). For every \( x, y \in V \) it holds that

\[
|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle. \quad (245)
\]

Furthermore, equality is achieved if and only if \( x \) and \( y \) are linearly dependent.

Proof. Note that we use this theorem to prove that the induced norm is a norm, so we cannot use the norm here. Associate with \( \langle \cdot, \cdot \rangle \) the function \( Q(x) := \langle x, x \rangle \). By lemma 764, \( Q \) is positive definite.

Fix \( x, y \in V \) and \( t \in \mathbb{C} \). If either vector is zero the statement is trivially true, so let both be nonzero. We have

\[
Q(x + ty) = \langle x + ty, x + ty \rangle = \\
= Q(x) + \bar{t} \langle x, y \rangle + t \langle y, x \rangle + |t|^2 Q(y) = \\
= Q(x) + 2 \text{real} \bar{t} \langle x, y \rangle + |t|^2 Q(y)
\]

Take \( t := -\frac{\langle x, y \rangle}{Q(y)} \), so that

\[
Q(x + ty) = Q(x) - 2 \frac{|\langle x, y \rangle|^2}{Q(y)} + \frac{|\langle x, y \rangle|^2}{Q(y)} = Q(x) - \frac{|\langle x, y \rangle|^2}{Q(y)}
\]

Since \( Q(x + ty) \geq 0 \), it follows that

\[
Q(x) - \frac{|\langle x, y \rangle|^2}{Q(y)} \geq 0
\]

\[
Q(x)Q(y) \geq |\langle x, y \rangle|^2.
\]

If \( x \) and \( y \) are linearly dependent, equality obviously holds. Conversely, suppose that equality holds. This implies that

\[
Q(x + ty) = 0,
\]

which by the positive definiteness of \( Q \) means that \( x = -ty \). Thus, \( x \) and \( y \) are linearly dependent. \( \square \)
**Definition 766.** Let $V$ be a real or complex inner product space with product $\langle \cdot, \cdot \rangle$. We define its induced norm as

$$ ||\cdot|| : V \rightarrow \mathbb{R}_{\geq 0} $$

$$ ||x|| := \sqrt{\langle x, x \rangle}. $$

If $V$ is a real inner product space, the induced norm is a square root of the induced quadratic form of $\langle \cdot, \cdot \rangle$.

**Proof.** We will only prove the complex case because the real case is identical, but slightly simpler.

Note that $||\cdot||$ is well-defined (that is, positive definite) by lemma 764.

Now we will show that it is a norm.

**Proof of N1.** Follows from the positive definiteness of $\langle \cdot, \cdot \rangle$

**Proof of N2.** For $t \in \mathbb{C}$ and $x \in V$ we have

$$ ||tx|| = \sqrt{\langle tx, tx \rangle} = |t| \sqrt{\langle x, x \rangle} = |t||x||. $$

**Proof of N3.** For $x, y \in V$ we have

$$ ||x + y||^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle $$

$$ = ||x||^2 + 2 \text{ real} \langle x, y \rangle + ||y||^2 \leq $$

$$ \leq ||x||^2 + 2|\text{real} \langle x, y \rangle| + ||y||^2 \overset{765}{=} $$

$$ = ||x||^2 + 2||x|| ||y|| + ||y||^2 = (||x|| + ||y||)^2 $$

Therefore,

$$ ||x + y|| \leq ||x|| + ||y||. $$

\[ \square \]
11.4. Algebraic dual spaces

**Definition 767.** Let \( V \) be a vector space over the field \( \mathbb{K} \). By [proposition 568](#), the set \( \text{hom}(V, \mathbb{K}) \) of all linear maps from \( V \) to the underlying field \( \mathbb{K} \) also form a vector space over \( \mathbb{K} \).

We call this space the **algebraic dual space** of \( V \) and denote it by \( V^* \). We call the functions in \( V^* \) **linear functionals**. The prefix “algebraic” is important when confusion is possible with **continuous dual spaces**.

**Remark 768.** The term “functional” as a noun has no definite meaning.

- In the context of linear algebra, and in particular definition 767, the term “functional” refers to “linear functional”, i.e. a linear map from a vector space to its base field. This terminology can be found, for example, in [Kna16, p. 50] and [Tyr04, sec. 26.1].
- In the context of functional analysis, “linear functional” may refer to either continuous linear functionals from some topological vector space to its base field, or to arbitrary linear functionals. This terminology can be found, for example, in [Rud91, def. 3.1] and [Cla13, sec. 1.3].

An arbitrary map from a topological vector space to its field may also be called a functional — for example, [KF80, p. 102] and [Dei85, p. 223] refer to “nonlinear functionals”. Minkowski functionals are notoriously nonlinear.

- In the context of recursive functions, for example in [Dea20], functionals are defined as “operations which map one or more functions of type \( \mathbb{N}^k \to \mathbb{N} \) (possibly of different arities) to other functions”.

The commonality between linear algebra and functional analysis is that “functional” refers to a map from a vector space to its base field. The commonality between functional analysis and logic is that “functional” refers to a map acting on a set of functions.

**Definition 769.** A **duality pairing** \( \langle \cdot, \cdot \rangle : U \times V \to \mathbb{K} \) is a bilinear function. such that, if \( \langle u, v \rangle = 0 \) for all \( u \), then \( v = 0 \).

**Definition 770.** Given a vector space \( V \), the following function is bilinear:

\[
\langle \cdot, \cdot \rangle : V^* \times X \to \mathbb{K} \\
\langle x^*, x \rangle \mapsto x^*(x).
\]

**Proposition 771.** Fix a vector space \( V \) over \( \mathbb{K} \) and a basis \( B \) of \( V \). For every basis vector \( e \), the **projection functional** \( \pi_e : V \to \mathbb{K} \), defined in definition 621, maps an arbitrary vector \( v \) to its \( e \)-th coordinate.

(a) Given basis vectors \( x \) and \( y \) from \( B \), \( \pi_x \) and \( \pi_y \) are linearly independent in \( V^* \).

(b) Furthermore, if \( e_1, \ldots, e_n \) is a basis of \( V \), then \( \pi_{e_1}, \ldots, \pi_{e_n} \) is a basis for the dual space \( V^* \).
(c) The set \( \{ \pi_e \mid e \in B \} \) spans \( V^* \) if and only if \( V \) is finite dimensional.

Proof.

**Proof of 771 (a).** Let \( t_x \) and \( t_y \) be scalars such that

\[ t_x \pi_x + t_y \pi_y = 0_V. \]

Then

\[ 0 = t_x \pi_x(x) + t_y \pi_y(x) = t_x \cdot 1 + t_y \cdot 0. \]

Analogously, \( t_y = 0 \). Therefore, the functionals \( \pi_x \) and \( \pi_y \) are linearly independent.

**Proof of 771 (b).** Let \( l \) be an arbitrary linear functional. Then

\[ l(y) \]

**Proof of 771 (c).**

**Proposition 772.** Fix a vector space \( V \). We define the canonical embedding into the double dual \( V^{**} \) of \( V \) by

\[
\Phi : V \rightarrow V^{**} \\
\Phi(x) := (\varphi \mapsto \varphi(x)),
\]

where \( \varphi \in V^* \).

**Proposition 773.** The dual vector space of a finite-dimensional vector space has the same dimension.

Proof. Let \( V \) be an \( n \)-dimensional vector space over \( F \) and let \( B \) be a basis of \( V \). For each \( b \in B \), define its dual vector on \( V^* \) as the linear extension of the functions

\[
\varphi : B \rightarrow F \\
\varphi(x) := \begin{cases} 
1, & x = b \\
0, & x \neq b
\end{cases}
\]

from the basis to the whole space. Denote the dual basis vector of \( b \) by \( b^* \).

We will now show that the set \( B^* := \{ b^* : b \in B \} \) forms a basis of \( V^* \).

Fix \( x^* \in V^* \). Define

\[
y^* := \sum_{b \in B} x^*(b) b^*.
\]

The linear functions \( x^* \) and \( y^* \) evidently agree on the basis \( B \). Hence, they agree on the whole space.

Hence, \( B^* \) is a basis of \( V^* \). Note that it has the same cardinality as the basis of \( B \).

**Remark 774.** By ?? ([UNDEFINED]), the vector space \( F^n \) is isomorphic to its dual \( F^{n^*} \).

In practice, it is sometimes useful to distinguish between vectors and functionals. This is why we regard functionals as either
functions

- column vectors
- row vectors

depending on what interpretation suits us best.

This is consistent with proposition 717, where we regard linear operators as matrices that act on vectors by multiplication.

For example, if we have the differentiable function \( f(x, y) = xy \), we can regard its gradient at the point \((x, y)\) as the row vector

\[
f'(x, y) = (y \ x).
\]

This is a linear functional that can acts on regular (column) vector by multiplying them from the left.

**Definition 775.** We define the dual linear operator of \( L : U \to V \) as

\[
L^* : V^* \to U^*
\]

\[
L^*(v^*) := v^*oL.
\]

**Definition 776.** Fix a subset \( S \subseteq V \) of a vector space \( V \) over \( F \). We define the annihilator of \( S \) as the vector space of functionals

\[
\text{ann}(S) := \{ x^* \in V^* : x^*(x) = 0_F \quad \forall x \in S \}.
\]

[Kna16] page 52
11.5. Diagonalization
12. Mathematical logic

Mathematical logic uses mathematics to study logic and vice versa.

We start with objects that are purely logical in nature — formulas — which are strings of symbols that represent truth values. Formal definitions for formulas are given here using grammars, which in turn depend on languages. Formal definitions for truth values are given using Heyting and Boolean algebras.

These definitions help us define the theory necessary to study two important intertwined topics:

- We are interested in establishing whether the formula $\varphi$ logically entails the formula $\psi$. This is done using deductive systems which specify precisely how we can manipulate strings of symbols. This aspect is called syntactic or logical and is the basis of proof theory. Formulas allow us to express statements about mathematics and proof theory allows us to systematically study the relationships between them.

- Given a formula, we are interested in assigning a meaning to it. Different logical systems provide different syntax that is useful for different purposes - propositional formulas allow us to express complex relationships between propositions via Boolean functions while first-order logic formulas allow us to go one level lower and give a precise meaning to these propositions via structures. This aspect of logic is called semantical and is the basis of model theory. Model theory allows us to study logical formulas using pre-existing mathematics.

There are two aspects in which logical systems are categorized:

- Propositional and first-order logic (among others) differ in what their syntax allows us to express. This also means that they differ in what their semantics can express, but, just as the syntax of first-order logic is a superset of the syntax of propositional logic, Boolean functions can express relations between quantifierless atomic formulas in any structures. In other words, semantics are identical in places where the syntax is the same.

- Classical and intuitionistic logic (among others) differ in their semantics and their logical inference rules. This has two implications
  - Boolean functions describe classical logic, however they fail to describe intuitionistic logic because double negation elimination (AX DNE) no longer holds and neither do other similar statements. So, while retaining the same syntax, we must resort to much more complicated semantical frameworks like Heyting or topological semantics.
  - The proof theory that describes classical logic no longer matches the semantics, hence we must resort to other proof systems. This turns out not to be trivial because we need a clear understanding of which logical axioms imply the others. Section 12.7 (Deductive systems) lists different proof systems and their corresponding semantics.
Remark 777. The statements of mathematical logic can themselves be studied logically. We distinguish between the object logic which we study and the metalogic which we use to study it. It is possible, for example, to study intuitionistic propositional logic using classical first-order logic. The metalogic is usually less formal and its statements are written in prose for the sake of easier understanding.

It is an exercise in futility to try and completely formalize the language, syntax and theory of the metalogic — the metalanguage, metasyntax and metatheory. We must take a given metalogical framework for granted and then study a certain object logical framework. This is not to say that the principles and rules that hold in the metalogic are immaterial — see for example the discussion of the differences between intuitionistic logic and classical logic. This is to say that it makes little sense to attempt to study the metalogic because at that point it becomes the object logic and the still more abstract conceptual framework in which we reason about the metalogic now becomes the new metalogic. We can thus form a hierarchy that is unbounded in both directions — we can study a more concrete object logic within the object logic, and we can jump from one metalogical level to the next.

An important connection between the logic and metalogic is given in remark 921.

Definition 778. Classical logic is a vague term that we use to describe a semantic framework, definition 813, and a matching deductive system, definition 898, for propositional logic and also a semantic framework, definition 832, and matching proof deduction deduction, definition 901, for first-order logic, among others.

It is characterized by the ability to use the law of double negation elimination (AX DNE). A more popular (but less accurate due to theorem 887) characterization is that the law of the excluded middle (AX LEM) holds. Within the metalogic, this law is called the principle of bivalence and states that either a statement holds or its negation holds.

Definition 779. Intuitionistic logic is a generalization of classical logic. It is also called constructive logic due to the Brouwer-Heyting-Kolmogorov interpretation. See remark 896 for further discussion of the topic.

Instead of the law of the excluded middle (AX LEM), we have the strictly weaker principle of explosion (AX EFQ) stating that everything can be proved given a contradiction.

To these ideas there correspond Heyting algebra semantics and topological semantics and a matching deductive system, definition 889, for propositional logic.

Definition 780. Minimal logic is a generalization of intuitionistic logic. Instead of the law of the excluded middle (AX LEM) and the strictly weaker principle of explosion (AX EFQ), we have the even weaker law of non-contradiction (AX LNC).

Metalogically speaking, we can only conclude that there is no statement such that both the statement and its negation are true. If the statement instead does not hold, we cannot automatically conclude that its negation holds.

Definition 883 provides a deductive system for propositional logic, but we avoid studying semantics of minimal logic.

Remark 781. We will only work in classical metalogic. Outside the section on logic, we will use formulas and, more generally, use object logic only in dedicated places like definition 455 (a)
describing the logical theory of groups. Most axioms like N1-N3 for norms are formulated entirely within the metalanguage under the assumption that we are working within a model of set theory. To keep a clear distinction between logical formulas and non-logical axioms and, more generally, to distinguishing between logic and metalogic, we use the following conventions:

(a) Variables in the object language are denoted by the small Greek letters, usually $\xi, \eta, \zeta$, while variables in the metalanguage are denoted by small Latin letters, usually $x, y, z$. If needed, we add subscripts with indices.

(b) Formulas, which we only consider in the object language, are also denoted by small Greek letters — $\varphi, \psi, \theta, \chi$ — and, so are terms — $\tau, \sigma, \rho, \kappa, \mu, \nu$.

(c) The propositional constants denoting truth and falsity are denoted by $\top$ and $\bot$ in the object language and by $T$ and $F$ in the metalanguage. This is only for the sake of following an established convention, and we still use $\top$ and $\bot$ in general lattices.

(d) We usually prefer prose to symbolic quantifiers and connectives in the metalanguage. The longer double arrows $\implies$ and $\iff$ are sometimes used within the metalogic outside this section.

(e) We conflate structures in the metalogic (i.e. sets with functions and/or relations defined on them) with their domain — see remark 834 for a discussion.

(f) We additionally use syntactic shorthands like remark 808 and remark 825 when writing formulas.

(g) We avoid writing excessive universal quantification and instead rely on implicit universal quantification as described in proposition 843. If we need the formulas to be closed, such as in the case of first-order theories for example, we assume all formulas are closed and if they are not, we add explicit universal quantifiers in front.

Some axioms like (159) are formulated within the metalogic for convenience and clarity, but are used as formulas in the object language in theorems like proposition 863. In places like this, it is usually straightforward to translate axioms from the metalogic into logical formulas.

Remark 782. Since we describe first-order logic, it may be helpful to clarify why is it named, so. It is merely a shorthand for “first-order predicate logic”. There are other predicate logical frameworks, namely second-order predicate logic (described in [Aut20, ch. VIII]) and higher-order predicate logic, also known as “simple type theory” (described in [Far08, sec. 3]).

Second-order logic allows us to quantify over relations between variables. In that case, we refer to the variables of first-order logic as “individuals” and to the relations as “relation variables”. This allows us, for example, to avoid axiom schemas like the axiom schema of specification by instead replacing them with a single axiom that quantifies over unary relations. A downside of second-order logic is that it has worse properties — it is incomplete.
in the sense that there exists no deductive system that is both, sound and complete\textsuperscript{8} and it is not compact in the sense that the analogue to theorem \textsuperscript{908} (First-order compactness theorem) does not hold\textsuperscript{9}. This is attributed to the expressive power of second-order logic because a first-order axiom schema may have only a countable number of axioms while a second-order quantifier may range over uncountably many relations.

Clearly anything that extends second-order logic must suffer from the same consequences, however higher-order logic is still useful because it allows us to utilize some very powerful concepts. Rather than quantifying over relations over the relations over individuals that would happen in third-order logic, we instead consider the more abstract frameworks of type theory. Type theory itself comes in many flavors, but simple type theory can be viewed as a generalization of first-order logic — see [Far08, thm. 2]. The rough idea is that rather than having individual variables, relation variables, etc., we have 
\textbf{base types} and \textbf{type constructors}. The individual variables have a dedicated base type, for example, and the types of functions and predicates are easily constructed using the basic type constructors, hence it is also easy to construct higher-order functions and predicates. The syntax of simple type theory is inspired by $\lambda$-calculus, which is a huge topic in itself and one of the frameworks for studying computability theory. The semantics of simple type theory are merely an extension of first-order semantics with different universes for different types. Like second-order logic, however, type theories have worse properties than those of first-order logic.

Another benefit of type theories is that they allow for multiple base types. For example, in the definition of a vector space, we have scalars and vectors, and we introduce an axiom schema parameterized by the scalars. In contrast, we could have a type for scalars and a type for vectors. This is also easily achievable in first-order logic via the, so-called “many-sorted first-order logic”, where the types are called “sorts”. We lack type constructors, and thus we are restricted in how our functions and predicates are defined, however for simple cases many-sorted first-order logic is just as useful as simple type theory. As a matter of fact, both many-sorted first-order logic languages and simple type theory languages can be reformulated as first-order logic languages — see [Far08, ch. 8].

We circumvent the need for any of these higher-order logical frameworks by using set theory — see \textbf{remark 922}. 

\textsuperscript{8} refer to [Aut20, thm. 39.6]
\textsuperscript{9} refer to [Aut20, thm. 39.7]
12.1. Formal languages

Languages are used to define formulas for expressing the axioms of set theory. Here, sets are used to formally define languages. A simple way out of this vicious cycle is via the theory-metatheory relationship discussed in remark 777 and remark 921. In short, we define languages within the metatheory using the already available concept of set, and we later define formulas, again in the metatheory, which allows us to subsequently formally define sets via axioms within the object logic.

Definition 783. Fix a nonempty set $\mathcal{A}$.

(a) We call $\mathcal{A}$ an alphabet.

(b) We call each element of $\mathcal{A}$ a symbol.

(c) A string over $\mathcal{A}$ is a finite sequence of symbols. If $(a, b, c)$ is a word, for convenience we write it as the string $abc$. This is the reason words are also referred to as strings. This notation only makes sense if each symbol of the language is actually represented by one typographic symbol.

The term “string” is common in programming practice. In the context of formal languages, strings are often called words. We will avoid the later term since it does not correspond to the everyday use of the term “word”.

(d) We denote the empty string by $\varepsilon$.

(e) The length $\text{len}(w)$ of a word $w$ is the number of elements of the tuple $w$.

(f) The concatenation of the words $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_m)$ is the word

$$vw := (v_1, \ldots, v_n, w_1, \ldots, w_m).$$

We abbreviate $vw \ldots w$ as $w^k$. This is only a notation. We do not distinguish, formally, between the words $aaabbaa$ and $a^3b^2a^2$, nor between $aeb$ and $ab$.

(g) The reverse word of $w = (w_1, \ldots, w_n)$ is

$$\text{rev}(w) := (w_n, \ldots, w_1).$$

(h) The word $p = (p_1, \ldots, p_m)$ is a prefix of $w = (w_1, \ldots, w_n)$ if

$$w = (p_1, \ldots, p_m, w_{m+1}, \ldots, w_n).$$

(i) The word $s$ is a suffix of $w$ if $\text{rev}(s)$ is a prefix of $\text{rev}(w)$. 

382
(j) The word \( v \) is a **subword** of \( w \) if there exists a prefix \( p \) and a suffix \( s \) of \( v \) such that
\[
    w = pvs.
\]

(k) The **Kleene star** \( \mathcal{A}^+ \) of \( \mathcal{A} \) is the set of all words over \( \mathcal{A} \). If we wish to exclude the empty word, like we often do, we instead write \( \mathcal{A}^+ \) for the set of all non-empty words over \( \mathcal{A} \). The Kleene star is also called the free monoid on \( A \) — see theorem 471 (Free monoid universal property).

(l) A **language** over \( \mathcal{A} \) is any subset of \( \mathcal{A}^* \). Note that in some contexts like propositional logic or first-order logic the term “language” may refer to the alphabet itself; see remark 804 for a further discussion.

**Definition 784.** Let \( V \) and \( \Sigma \) be disjoint nonempty subsets of some alphabet.

(a) We call elements of \( \Sigma \) **terminals**. We denote terminals in abstract grammars using lowercase Greek letters, and we denote words using lowercase Latin letters.

(b) We call elements of \( V \) **non-terminals**. By convention, variables are denoted using capital letters.

(c) We assume that a special **start symbol** \( S \in V \) is fixed.

(d) We define a binary relation \( \rightarrow \) of **production rules** over \( (V \cup \Sigma)^* \).

   We impose the restriction that no rules of the form \( \varepsilon \rightarrow v \) exist for any word \( v \). We do allow, however, production rules of the form \( v \rightarrow \varepsilon \). Such rules are called **\( \varepsilon \)-rules**.

   Rules describe transformations that define how a language is “generated” starting from \( S \). See definition 788 and example 789.

(e) The quadruple \( G := (V, S, \Sigma, \rightarrow) \) is called a **formal grammar** or simply a **grammar**.

**Definition 785.** We will define the **Chomsky hierarchy** of formal grammars. We can classify a grammar \( G = (V, S, \Sigma, \rightarrow) \) as follows, based on their rules:

(a) In general, every production rule has the form \( v \rightarrow w \), where both \( v \) and \( w \) are words consisting of terminal and non-terminals, and \( v \) is nonempty.

   When no additional restrictions are imposed on the rules of the grammar, we call it **unrestricted grammar**. The other levels of the hierarchy are subsets of the unrestricted grammars.

(b) The grammar is **non-contracting** if \( \text{len}(v) \leq \text{len}(w) \) for every production rule \( v \rightarrow w \).

(c) The grammar is **context-sensitive** if every rule has the form \( aAb \rightarrow w \) for some non-terminal \( A \), arbitrary words \( a \) and \( b \), and a nonempty word \( w \).

   The requirement for \( w \) to be nonempty is set up so that context-sensitive grammars are non-contracting.
(d) The grammar is **context-free** every rule has the form $A \rightarrow w$ for some non-terminal $A$ and a nonempty word $w$.

Unlike for context-sensitive languages, $w$ is allowed to be empty. Thus, a context-free grammar is non-contracting, however it may not be context-sensitive if it has $\varepsilon$-rules.

(e) Finally, the grammar is **regular** if every rule has one of the forms

$$
A \rightarrow \varepsilon, \\
A \rightarrow B\tau, \\
A \rightarrow \tau B, \\
A \rightarrow \tau,
$$

where $A$ and $B$ are non-terminals and $\tau$ is a terminal.

Regular grammars are obviously context-free.

**Example 786.** We define a grammar for primary school notation of multiplication and division of natural numbers. Note that we consider the numbers in $\mathbb{N}$ only as symbols, without any regard to semantics.

Let $V := \{N, O, M, E\}$ and $\Sigma := \mathbb{N} \cup \{\times, \div, (, )\}$. Define the grammar

$$
N \rightarrow 0 \\
N \rightarrow 1 \\
\vdots \\
N \rightarrow n \\
\vdots \\
O \rightarrow \times \\
O \rightarrow \div \\
E \rightarrow N \\
E \rightarrow (EOE)
$$

It is convenient to use the following shorthands:

$$
N \rightarrow 0 \mid 1 \mid 2 \mid \ldots \\
O \rightarrow \times \mid \div \\
E \rightarrow N \mid (EOE)
$$

We can choose different non-terminals as the starting symbol. The symbol $N$ corresponds to numbers, $O$ corresponds to operations, and $E$ can be either a number or an expression. We say that (247) specifies a grammar schema. With any starting symbol, the grammar is clearly context-free.

**Remark 787.** The infinitude of possible rules in example 786 may not bother us formally, but when dealing with software implementations, we must have a finite number of rules. An
example of a nontrivial grammar in the wild is the Python grammar that can be found in [Pyt21]. There are also other advantages of introducing a more convenient metasyntax (a syntax for describing language syntax).

For context-free grammars, it is often convenient to use the Backus-Naur form (BNF). For example 786, the BNF is

\[
\begin{align*}
\langle \text{nonzero digit} \rangle & ::= 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9 \\
\langle \text{digit} \rangle & ::= 0 \mid \langle \text{nonzero digit} \rangle \\
\langle \text{number} \rangle & ::= \langle \text{nonzero digit} \rangle \mid \langle \text{number} \rangle \langle \text{digit} \rangle \\
\langle \text{operation} \rangle & ::= \times \mid \div \\
\langle \text{expression} \rangle & ::= \langle \text{number} \rangle \mid (\langle \text{number} \rangle \langle \text{operation} \rangle \langle \text{number} \rangle ) .
\end{align*}
\]

The obvious difference is that we explicitly define numbers via their decimal representation, which means that we get a finite amount of rules. Compared to (246) some other differences are:

- Variables are denoted by \( \langle \text{words enclosed in angle brackets} \rangle \), so that we can name variables more descriptively using more than one symbol.
- Terminals are, by convention, put in “quotes”. In human-readable rich text documents like this one, it is sometimes possible to use different fonts, and so instead of using “quotes” we specify terminals using an upright typewriter font.
- Free-text rules can be specified using a normal font. This is also only used in human-readable rich text documents, however this usage is justified because such rules are only beneficial for human understanding and not for machine parsing.
- By convention, the symbol \( ::= \) is used instead of \( \rightarrow \) for specifying transition rules.
- Different rules with the same, source are concatenated as in (247).
- In order to fully describe a context-free grammar, we must only specify its Backus-Naur form and its starting variable.

**Definition 788.** Fix a formal grammar \( G = (V, S, \Sigma, \rightarrow) \).

(a) We define the binary relation \( \Rightarrow \) on the Kleene star \( (V \cup \Sigma)^* \) by declaring that, for every two words \( p \) and \( s \) over \( V \cup \Sigma \) and every production rule \( v \rightarrow w \), we have \( pw \Rightarrow ps \). We also define the relation \( \Rightarrow_L \), as a restriction of \( \Rightarrow \) to the cases where \( p \) contains only terminal symbols and \( \Rightarrow_R \) — if \( s \) contains only terminal symbols.

A derivation of the word \( w_n \) from \( w_1 \) is a directed path in the quiver induced by the relation \( \Rightarrow \), i.e.

\[
\begin{align*}
w_1 \Rightarrow w_2 \Rightarrow \cdots \Rightarrow & \cdots \Rightarrow w_{n-1} \Rightarrow w_n .
\end{align*}
\] (248)

A leftmost derivation is a derivation performed using \( \Rightarrow_L \) rather than \( \Rightarrow \). Rightmost derivations are defined analogously.

385
We say that $w_n$ is **derivable** from $w_1$ if there exists a derivation from $w_1$ to $w_n$.

We denote the **transitive closure** of \( \Rightarrow \) by \( \Rightarrow^+ \) and the **reflexive closure** of \( \Rightarrow \) by \( \Rightarrow^* \). Clearly $w_1$ is derivable from $w_n$ if and only if $w_1 \Rightarrow^* w_n$.

The leftmost and rightmost derivations generate the same derivability relation — the only potential difference is in the order of rule applications in the derivation itself.

**Example 789.** We continue example 786. Depending on our choice of starting symbol, we can derive different sets of words.

For the sake of simplifying our exposition and proof, however, we will assume the simpler grammar described in (246).

Choose the starting symbol to be $E$. We will show that this grammar is unambiguous. **Figure 24** demonstrates that removing the parentheses makes even this simple grammar ambiguous.

We will show that $G$ is unambiguous. Let $w$ be a word in $\mathcal{L}(G)$. We explicitly build the leftmost derivation of $w$ using recursion on $\text{len}(w)$:

- If $\text{len}(w) = 1$, then $w = n \in \mathbb{N}$, and the word has been derived as $E \rightarrow N \rightarrow n$.
- Assume that $w$ is unambiguously derived for $\text{len}(w) < m + 2$ and let $\text{len}(w) = m + 2$, then $w$ is necessarily enclosed in parentheses. Let $w = (\sigma_1 \ldots \sigma_m)$ be the symbols of $w$. Because of the parentheses, the only possibility for $\sigma_1 \ldots \sigma_m$ is that it consists of two
words in $\mathcal{L}(G)$ with either a multiplication symbol $\times$ or a division symbol $\div$ between them. Let $k$ be the index of the operator symbol, that is, the index such that $\sigma_1 \ldots \sigma_{k-1}$ and $\sigma_{k+1} \ldots \sigma_m$ both belong to $\mathcal{L}(G)$.

By the inductive hypothesis, both $\sigma_1 \ldots \sigma_{k-1}$ and $\sigma_{k+1} \ldots \sigma_m$ are unambiguously derived from $E$. Then $w$ is generated by the rule $E \rightarrow (EOE)$, where the operator symbol $\sigma_k$ determines the terminal of $O$. Therefore, the derivation of $w$ is also unambiguous.

**Definition 790.** An **ordered arborescence** is an arborescence $T = (G, A)$ with a partial order $\leq$ such that every set of siblings is a chain.

**Definition 791.** Fix a context-free formal grammar $G = (V, S, \Sigma, \rightarrow)$.

For every word $w$ in $\mathcal{L}(G)$, we will build an **ordered rooted tree** whose **leaves** are the symbols of $w$ and whose root is $S$. We will call this a **concrete syntax tree** for $w$.

Fix a derivation

$$S \Rightarrow w_1 \Rightarrow \cdots \Rightarrow w_{n-1} \Rightarrow w_n.$$  \hfill (249)

In practice, we want this to be unique and hence we can restrict ourselves to leftmost derivations in **unambiguous** grammars.

We use **natural number recursion** on $n$ to build the tree. Note that, for the purposes of recursion, we allow $w_n$ to contain non-terminals.

- In the trivial case where $n = 0$, and there is no actual derivation, we build a single-vertex tree with root $S$.

- Suppose that we can build a tree for all derivations of length $m - 1$ and fix a derivation (249) of length $n$.

First, build a tree $T$ from the derivation

$$S \Rightarrow w_1 \Rightarrow \cdots \Rightarrow w_{n-2} \Rightarrow w_{n-1}.$$
Figure 25: A concrete syntax tree for the expression \((6 \div (3 \times 2))\) from fig. 23

There must exist words \(p, s\) and \(v\) and a non-terminal \(A\) such that

\[ w_{n-1} = pAs \Rightarrow pvs = w_n. \]

There already exists a leaf for \(A\) in \(T\). For every symbol in \(v\), add a new node as a child of this node.

**Theorem 792** (Structural induction on unambiguous grammars). Unlike for the other induction principles in remark 1027, we will not formulate this one via logical formulas. This will complicate us unnecessarily. We will instead describe how the principle is used. It should be noted that, in applications of this principle, we prefer using abstract syntax trees to concrete syntax trees, but formulating the principle itself is easier via concrete syntax trees.

Let \(G = (V, S, \Sigma, \rightarrow)\) be an unambiguous context-free formal grammar. Suppose that we want to prove a statement for every word in \(L(G)\). It is sufficient to perform the following for every rule \(A \rightarrow w\):

Let \(A_1, ..., A_n\) be all non-terminals of \(w\) and let \(u_0, ..., u_n\) be subwords of \(w\) such that

\[ w = u_0A_1u_1A_2 ... A_nu_n. \]

Let \(v_1, ..., v_n\) be arbitrary words in \(L(G)\) derivable from \(A_1, ..., A_n\), respectively, so
Figure 26: An abstract syntax tree for the expression \( (6 \div (3 \times 2)) \) from fig. 23

that we have the concrete syntax tree

Then we must prove the statement for the word

\( u_0v_1u_1 \ldots v_nu_n. \)

Compare this principle to the more general theorem 997 (Well-founded induction).

Proof. Clearly every word in \( \mathcal{L}(G) \) can be obtained in this way. The principle itself follows from theorem 997 (Well-founded induction) even without non-ambiguity.

Non-ambiguity ensures that every word has a unique concrete syntax tree, and prevents us from proving the statement for one tree of a word and disproving it for another.

Remark 793. Unlike concrete syntax trees, which focus on how words are built from symbols, abstract syntax trees focus on how words are built with respect to semantics. For example, natural number arithmetic is not concerned with the symbols themselves but only in how operations are applied to numbers. This is better expressed via the abstract syntax tree in fig. 26 rather than the concrete syntax tree in fig. 25.

A formal definition for an abstract syntax tree would be clunky. It is an ordered rooted tree of words that unambiguously encodes a concrete syntax tree. The structure of an abstract syntax tree is determined entirely by the concrete application, and we will not attempt to give a more precise definition.

Applications of theorem 792 (Structural induction on unambiguous grammars) use abstract syntax trees as can be seen in section 12.3 (Propositional logic) and section 12.4 (First-order logic), and a broader discussion on the properties of certain trees can be found in remark 794.
Remark 794. Binary operations are easily extended to higher arities. Given a binary operation \( + : M \times M \to M \), we can extend it via natural number recursion to arbitrary \( n \)-tuples \( x_1, \ldots, x_n \) as

\[
(x_1 + (x_2 + \cdots + (x_{n-1} + x_n) \cdots)).
\]

This expression corresponds to the abstract syntax tree

\[
\begin{aligned}
 & x_1 \\
 & \quad + \\
 & x_2 \\
 & \quad + \\
 & \quad \vdots \\
 & x_{n-2} \\
 & \quad + \\
 & x_{n-1} \\
 & \quad + \\
 & x_n
\end{aligned}
\]

Several things can be noted here.

(a) When exchanging the order of the parentheses in the expression (250), only the root is changed in the syntax tree (251). Therefore, for an associative binary operation, abstract syntax trees can instead be represented as finite ordered rooted trees like

\[
\begin{aligned}
 & x_1 \\
 & \quad + \\
 & x_2 \\
 & \quad \vdots \\
 & x_{n-2} \\
 & \quad + \\
 & x_{n-1} \\
 & \quad + \\
 & x_n
\end{aligned}
\]

Of course, \( n \)-ary trees can still be used for non-associative binary operations, as long as we have selected a strategy for evaluation. If we evaluate (252) as (250), we say that the operation is right associative. Dually, if we evaluate (252) as

\[
((\cdots (x_1 + x_2) + \cdots + x_{n-1}) + x_n),
\]

we say that the operation is left associative.

Note that, unlike associativity, left and right associativity are not properties of the operation, but rather conventions on how to evaluate expressions without explicit parentheses. For example, in a Heyting algebra, the conditional \( \to \) is not associative, but it is often taken to be right associative so that

\[
x \to y \to z
\]

is evaluated as

\[
(x \to (y \to z)).
\]

(b) If, additionally, the operation is commutative, we can regard the syntax tree as unordered.

Note that extending operations can be confusing for commutative but non-associative operations. Commutativity allows us to swap summands inside parentheses, and associativity is needed to “remove” the parentheses.
Consider the real number midpoint operation
\[ x \oplus y := \frac{x + y}{2} \]
from example 442 (c). It is commutative and not associative, and in the expression
\[ x \oplus y \oplus z \]
we can swap \( x \) with \( y \) but not with \( z \). This can be very unintuitive. We aim
to always write parentheses for non-associative operations.

(c) Suppose that \(+\) is associative and commutative. Suppose also that we are given an
indexed family \( \{x_k\}_{k \in \mathcal{K}} \) of elements of \( M \). We can obviously construct a tree with root
+ and children \( \{x_k\}_{k \in \mathcal{K}} \).

This is not strictly a syntax tree. Syntax trees are necessarily finite, and the family may
even be uncountable. Nevertheless, we can sometimes evaluate this tree to obtain a
member of the monoid.

(i) The operations of a lattice arise by specializing suprema and infima to binary
sets. Conversely, as discussed in proposition 1255 (e), the binary lattice operations
induce a partial order, and we can define suprema and infima. If the lattice
happens to be complete, instead of the binary operations, we can use
\[ \bigvee_{k \in \mathcal{K}} x_k = \text{sup}\{x_k \mid k \in \mathcal{K}\}. \]

(ii) If \( M \) is a topological monoid and if \( \mathcal{K} = \{1, 2, 3, \ldots\} \), the family is a sequence, and
we can define the sequence of partial sums
\[ \left\{ \sum_{k=1}^{n} x_k \right\}_{n=1}^{\infty} \]
This gives rise to series discussed in section 3.2 (Series) and section 2.5 (Real
series). A limit may not exist for the net, unfortunately, and if it does, it may not
be unique (if the topology is not Hausdorff).

(iii) Suppose that \( M \) has an identity element \( 0_M \). If only finitely many elements of
the family are different from \( 0_M \), we regard the ordinary summation operation as
well-defined on the whole family and write
\[ s := \sum_{k \in \mathcal{K}} x_k. \]

Technically, this involves selecting a well-ordering \( x_{1_n}, \ldots, x_{k_n} \) on the set
\[ \{x_k \mid k \in \mathcal{K} \text{ and } x_k \neq 0\} \]
and assigning to \( s \) the result of the iterated binary operation
\[ (x_{1_n} + (x_{2_n} + \cdots + (x_{k_{n-1}} + x_{k_n})\cdots)). \]
Commutativity ensures that the sum $s$ does not depend on the order of summands (and hence on the well-order we have chosen). Adding any member of the family would not change the sum, which justifies this shorthand definition. Furthermore, since we only sum finitely many summands, we can construct a well-ordering using natural number recursion without relying on the axiom of choice.

This is fundamental for the definition of direct sums, which in turn are used to define linear combinations and polynomials.
12.2. Boolean functions

**Definition 795.** Fix a two-element set \( \{ T, F \} \). We can think of \( T \) as a value denoting truth and \( F \) as denoting falsity. See remark 781 (c) for notation conventions.

There is a natural **Boolean algebra** structure on \( \{ T, F \} \) where \( T \) is the top and \( F \) is the bottom and the operations are defined in an obvious way.

By proposition 1276, all two-element Boolean algebras are isomorphic, so our choice of symbols does not matter much. We do sometimes identify \( T \) with 1 and \( F \) with 0 in the prime field \( \mathbb{F}_2 \). The latter is a Boolean algebra as discussed in corollary 1277.

**Definition 796.** We call functions from any set to \( \{ T, F \} \) (Boolean-valued) **predicates** and functions from \( \{ T, F \}^n \) to \( \{ T, F \} \) **Boolean functions**.

**Remark 797.** Boolean-valued functions and relations represent the same concept. In particular, the relation \( R \subseteq A_1 \times \cdots \times A_n \) corresponds to a unique Boolean-valued function

\[
 f : X_1 \times \cdots \times X_n \to \{ T, F \}
\]

\[
 f(x_1, ... , x_n) = \begin{cases} 
 T, & (x_1, ..., x_n) \in R, \\
 F, & \text{otherwise} 
\end{cases}
\]

and vice versa.

**Definition 798.** Fix a set \( B \) of Boolean functions of arbitrary arities.

The **closure** \( \text{cl} B \) of \( B \) is defined recursively as follows:

- If \( f \in B \), then \( f \in \text{cl} B \)
- If \( f_k(x_1, ..., x_n) \in \text{cl} B \) for \( k = 1, ..., m \) and if \( g(x_1, ..., x_m) \in \text{cl} B \), then their superposition

\[
 h(x_1, ..., x_n) := g(f_1(x_1, ..., x_n), ..., f_m(x_1, ..., x_n))
\]

is also in \( \text{cl} B \).

We say that \( B \) is **closed** if \( \text{cl} B = B \) and **complete** if \( \text{cl} B \) is the set of all Boolean functions of arbitrary arity.

If \( B \) is complete, then from proposition 866 it follows that \( B \) is a Boolean algebra. This is used in theorem 814 (b).

**Definition 799.** A **Zhegalkin polynomial** is a polynomial in the prime field \( \mathbb{F}_2 \). Due to proposition 699, however, we restrict ourselves to polynomials with square-free monomials.

For example, for every binary Boolean function there exist coefficients \( a, b, c, d \in \mathbb{F}_2 \) such that

\[
 f(x, y) = axy \oplus bx \oplus cy \oplus d. \tag{253}
\]

**Definition 800.** Unlike arbitrary functions, Boolean functions only have a small finite number of possible values that can easily be enumerated.

Out of the following binary operations, \( \lor \), \( \land \) and \( \neg \) form the **Boolean algebra** structure on \( \mathbb{F}_2 \) and \( \oplus \) and \( \land \) form the **field** structure on \( \mathbb{F}_2 \). The operations \( \rightarrow \) and \( \leftrightarrow \) are also defined in any **Boolean algebra**.
See proposition 815 for direct consequences of these definitions.

**Definition 801.** Fix a Boolean function $f(x_1, \ldots, x_n)$ in the prime field $\mathbb{F}_2$.

(a) Its **dual function** is

$$\bar{f}(x_1, \ldots, x_n) := f(\overline{x_1}, \ldots, \overline{x_n}).$$

(b) $f$ is **self-dual** if it is its own dual.

(c) $f$ is **truth-preserving** if $f(T, \ldots, T) = T$.

(d) $f$ is **falsity-preserving** if $f(F, \ldots, F) = F$.

(e) $f$ is **monotone** if, for any two tuples of arguments $x_1, \ldots, x_n \in \mathbb{F}_2$ and $y_1, \ldots, y_n \in \mathbb{F}_2$, the inequalities $x_k \leq y_k$ for all $k \in \{1, \ldots, n\}$ imply that

$$f(x_1, \ldots, x_n) \leq f(y_1, \ldots, y_n).$$

(f) $f$ is **linear** if its Zhegalkin polynomial is linear, i.e. has only monomials of degree 0 or 1. In the case of binary Boolean functions, this means that the coefficient $a$ in (253) is zero.

**Theorem 802** (Post-Yablonsky completeness theorem). The family $B$ of Boolean functions is complete if and only if all of the following conditions are satisfied:

(a) $B$ contains a function that is not truth-preserving.

(b) $B$ contains a function that is not falsity-preserving.

(c) $B$ contains a function that is not self-dual.

(d) $B$ contains a function that is not monotone.

(e) $B$ contains a function that is not linear.

**Example 803.** We give examples of complete sets of Boolean functions in $\mathbb{F}_2$.

(a) The archetypic example of a complete set of Boolean functions is the triple $\lor, \land, \top$ that forms the Boolean algebra structure on $\mathbb{F}_2$.

We verify that the conditions of **Theorem 802** (Post-Yablonsky completeness theorem) are satisfied:
802 (a) \( \overline{\top} \) is not truth-preserving.

802 (b) \( \overline{\top} \) is not falsity-preserving.

802 (c) Neither \( \lor \) nor \( \land \) are self-dual. In fact, due to theorem 1279 (De Morgan’s laws), \( \land \) is the dual of \( \lor \) and vice versa.

802 (d) \( \overline{\top} \) is not monotone.

802 (e) Neither \( \lor \) nor \( \land \) have linear Zhegalkin polynomials.

Thus, \( \{\land, \lor, \overline{\top}\} \) is indeed a complete set of Boolean functions. Note that having both \( \lor \) and \( \land \) is redundant and we usually include both for symmetry. The families \( \{\land, \overline{\top}\} \) and \( \{\lor, \overline{\top}\} \) are both complete.

This is utilized for conjunctive and disjunctive normal forms.

(b) We can go even further and have a single binary Boolean function generate all others. We will use the function

\[
(x \uparrow y) := \overline{x} \land y = xy \oplus 1.
\]  

(254)

This operation is called Sheffer’s stroke or nand (“not and”).

We have

\[
\overline{x} = (x \uparrow 1) \quad \text{and} \quad (x \land y) = \overline{x} \uparrow \overline{y},
\]

which allows us to reduce the case to example 803 (a). We conclude that the singleton set \( \{1\} \) is a complete set of Boolean operations.

(c) Another commonly used complete family is \( \{\rightarrow, \overline{\top}\} \). We verify that the conditions of theorem 802 (Post-Yablonsky completeness theorem) are satisfied:

802 (a) \( \overline{\top} \) is not truth-preserving.

802 (b) \( \rightarrow \) is not falsity-preserving because \( (F \rightarrow F) = T \).

802 (c) \( \rightarrow \) is not self-dual because \( \overline{x} \rightarrow \overline{y} \overset{(262)}{=} (y \rightarrow x) \neq (x \rightarrow y) \).

802 (d) \( \rightarrow \) is not monotone because \( F \rightarrow T = F \).

802 (e) \( \rightarrow \) doesn’t have a linear Zhegalkin polynomial.

(d) Given the family \( \{\rightarrow, F\} \), we can define

\[
\overline{x} := (x \rightarrow F),
\]

which shows that \( \{\rightarrow, F\} \) is also a complete family.
12.3. Propositional logic

*Remark* 804. The language of propositional logic is, strictly speaking, an alphabet rather than a language. Nonetheless, this is the established terminology.

*Definition 805.* The language of propositional logic consists of:

(a) A nonempty, at most countable set **Prop** of propositional variables. Technically, we can have different languages with different variables, but it is safe to assume that there is only one single language of propositional language.

(b) Two propositional constants (also known as truth values):
   - The **verum** \(\top\).
   - The **falsum** \(\bot\).

(c) Negation \(\neg\).

(d) The set \(\Sigma\) of propositional connectives, namely
   - **Conjunction** \(\land\) (also known as and and meet).
   - **Disjunction** \(\lor\) (also known as or and join).
   - **Conditional** \(\rightarrow\) (also known as if...then and material implication).
   - **Biconditional** \(\iff\) (also known as iff and material equivalence).

   Note that “conditional” and “biconditional” are nouns in this context.

(e) Parentheses ( and ) for defining the order of operations unambiguously (see *remark 808* for a further discussion).

*Remark 818* shows we can actually utilize a smaller propositional language without losing any of its semantics.

*Definition 806.* The following related definitions constitute what is called the syntax of propositional logic.

(a) Consider the following grammar schema:

\[
\begin{align*}
\langle \text{variable} \rangle & ::= P \in \text{Prop} \\
\langle \text{connective} \rangle & ::= \circ \in \Sigma \\
\langle \text{formula} \rangle & ::= \langle \text{variable} \rangle \mid \top \mid \bot \mid \neg \langle \text{formula} \rangle \mid (\langle \text{formula} \rangle \langle \text{connective} \rangle \langle \text{formula} \rangle )
\end{align*}
\]

Note that **Prop** may be infinite, in which case the grammars may have infinitely many rules. If needed, we can circumvent this by introducing an appropriate naming
convention for variables, for example by allowing arbitrary strings of alphanumeric characters for variable names.

For the sake of readability, we will be using the conventions in remark 808 regarding parentheses.

(b) The set Form of propositional formulas is the language generated by this grammar schema with (formula) as a starting rule. Propositional formulas are also called sentences unlike in first-order logic where only specific formulas are called sentences — see definition 823 (k).

The grammar of propositional formulas is unambiguous as shown by proposition 807, which makes it possible to perform proofs via theorem 792 (Structural induction on unambiguous grammars).

(c) If \( \varphi \) and \( \psi \) are formulas and \( \psi \) is a subword of \( \varphi \), we say that \( \psi \) is a subformula of \( \varphi \).

(d) For each formula \( \varphi \), we inductively define its variables to be elements of the set

\[
\text{Var}(\varphi) := \begin{cases}
\emptyset, & \varphi \in \{\top, \bot\} \\
\{P\}, & \varphi = P \in \text{Prop} \\
\text{Var}(\psi), & \varphi = \neg \psi \\
\text{Var}(\psi) \cup \text{Var}(\theta), & \varphi = \psi \circ \theta, \circ \in \Sigma.
\end{cases}
\]  

Note that \( \text{Var}(\varphi) \) can naturally be totally ordered by the position of the first occurrence of a variable.

**Proposition 807.** The grammar of propositional formulas is unambiguous.

**Proof.** The proof is analogous to example 789.

**Remark 808.** We use the following two “abuse-of-notation” conventions regarding parentheses:

(a) We may skip the outermost parentheses in formulas with top-level connectives, e.g. we may write \( P \land Q \) rather than \( (P \land Q) \).

(b) Because of the associativity of \( \land \) and \( \lor \), which is implied by definition 812 and definition 800, we may skip the parentheses in chains like

\[
(...((P_1 \land P_2) \land P_3) \land ... \land P_{n-1}) \land P_n.
\]

and instead write

\[
P_1 \land P_2 \land ... \land P_{n-1} \land P_n.
\]

(c) Although not formally necessary, for the sake of readability we may choose to add parentheses around certain formulas like

\[
\neg P \lor \neg Q.
\]
and instead write

\((\neg P) \lor \neg Q\).

This latter convention is more useful for quantifiers in first-order formulas.

These are only notations shortcuts in the metalanguage and the formulas themselves (as abstract mathematical objects) are still assumed to contain parentheses that help them avoid syntactic ambiguity.

**Definition 809.** Theorems in mathematics usually have the form \(P \rightarrow Q\). Formulas of this form are called **material implications** in order to distinguish them from logical implication, which relates to the metatheoretic concept of entailment. This is further discussed in [Mum11]. Note that the term “material implication” sometimes also refers to the **conditional connective** \(\rightarrow\) itself.

We introduce terminology that is conventionally used when dealing with theorems.

(a) \(P\) is a **sufficient condition** for \(Q\).
(b) \(Q\) is a **necessary condition** for \(P\).
(c) \(P\) the **antecedent** of \(\varphi\).
(d) \(Q\) the **consequent** of \(\varphi\).
(e) The formula \(\neg P \rightarrow \neg Q\) is the **inverse**\(^{10}\) of \(\varphi\).
(f) The formula \(Q \rightarrow P\) is the **converse**\(^{11}\) of \(\varphi\).
(g) The formula \(\neg Q \rightarrow \neg P\) is the **contrapositive** of \(\varphi\). In classical logic, it is **equivalent** to the original formula due to **proposition 815 (c)**.

**Definition 810.** We define **valuations**\(^{12}\) for propositional formulas. It is possible to define different valuations, so in case of doubt, we will refer to the one defined here as the **classical valuation** giving classical semantics.

This valuation implicitly depends on the Boolean algebra fixed in **definition 796**. When dealing with **Heyting semantics**, we use more general Heyting algebras where not only the top and bottom, but also other values are utilized.

(a) A **propositional interpretation** is a function with signature \(I : \text{Prop} \rightarrow \{T,F\}\). See **definition 795** for remarks regarding the **Boolean algebra** \(\{T,F\}\) and the **definition 800** for a list of some standard Boolean operators.

---

\(^{10}\)bg: противоположна, ru: противоположная

\(^{11}\)bg: обратна, ru: обратная

\(^{12}\)bg: оценка, ru: оценка
Given an interpretation \( I \), we define the valuation of a formula \( \varphi \) inductively as

\[
\varphi[I] := \begin{cases} 
T, & \varphi = \top \\
F, & \varphi = \bot \\
I(P), & \varphi = P \in \text{Prop} \\
\psi[I], & \varphi = \neg \psi \\
\psi_1[I] \circ \psi_2[I], & \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma,
\end{cases}
\]

where \( \circ \) on the left denotes the Boolean operator corresponding to the connective \( \circ \) on the right.

Remark 811. If we know that \( \text{Var}(\varphi) \subseteq \{P_1, \ldots, P_n\} \), it follows that the valuation \( \varphi[I] \) only depends on the particular values \( I(P_1), \ldots, I(P_n) \) of \( I \).

Let \( x_1, \ldots, x_n \in \{F, T\} \) and let \( I \) be such that \( I(P_k) = x_k \) for \( k = 1, \ldots, n \). We introduce the notation

\[
\varphi[x_1, \ldots, x_n]
\]

for \( \varphi[I] \) because the rest of the interpretation \( I \) plays no role here. We may also use

\[
\varphi[\psi_1, \ldots, \psi_n]
\]

to denote substitution.

When using this notation, we implicitly assume that \( \text{Var}(\varphi) \subseteq \{P_1, \ldots, P_n\} \).

Definition 812. Let \( \varphi \) be a propositional formula and let \( \text{Var}(\varphi) = \{P_1, \ldots, P_n\} \) be an ordering of the free variables of \( \varphi \). We define the Boolean function

\[
\text{fun}_\varphi : \{T, F\}^n \to \{T, F\} \\
\text{fun}_\varphi(x_1, \ldots, x_n) := \varphi[x_1, \ldots, x_n].
\]

Definition 813. We now define semantical properties of propositional formulas. Because of the connection with Boolean functions given in Definition 812, we also formulate some of the properties using Boolean functions.

(a) Given an interpretation \( I \) and a set \( \Gamma \) of formulas, we say that \( I \) satisfies \( \Gamma \) if, for every formula \( \varphi \in \Gamma \) we have \( \varphi[I] = T \).

We also say that \( I \) is a model of \( \Gamma \) and write \( I \models \Gamma \).

If \( \Gamma = \{\gamma_1, \ldots, \gamma_n\} \) is a finite ordered set, we use the shorthand \( I \models \gamma_1, \ldots, \gamma_n \) rather than \( I \models \{\gamma_1, \ldots, \gamma_n\} \). In particular, if \( \Gamma = \{\varphi\} \) we write \( I \models \varphi \).

Note that every interpretation vacuously satisfies the empty set \( \Gamma = \emptyset \) of formulas.

We say that \( \Gamma \) is satisfiable if there exists a model for \( \Gamma \).

(b) We say that the set of formulas \( \Gamma \) entails the set of formulas \( \Delta \) and write \( \Gamma \models \Delta \) if either of the following hold:
Every model of $\Gamma$ is also a model of $\Delta$.

The following preimage inclusion holds:

$$\bigcap_{\varphi \in \Gamma} \text{fun}_\varphi^{-1}(T) \subseteq \bigcap_{\psi \in \Delta} \text{fun}_\psi^{-1}(T).$$

(c) The formula $\varphi$ is a (semantic) tautology if either:

- Every interpretation satisfies $\varphi$.
- The empty set $\Gamma = \emptyset$ of formulas entails $\varphi$, i.e. $\models \varphi$.
- The function $\text{fun}_\varphi$ is canonically true.

We also say that $\varphi$ is valid.

(d) Dually, $\varphi$ is a (semantic) contradiction if either:

- No interpretation satisfies $\varphi$.
- The formula $\varphi$ entails $\bot$, i.e. $\varphi \models \bot$.
- The function $\text{fun}_\varphi$ is canonically false.

(e) We say that $\varphi$ and $\psi$ are semantically equivalent and write $\varphi \equiv \psi$ if either:

- We have $\varphi[I] = \psi[J]$ for every interpretation $I$.
- Both $\varphi \models \psi$ and $\psi \models \varphi$.

(f) A weaker notion than that of semantic equivalence is that of equisatisfiability. We say that the families $\Gamma$ and $\Delta$ are equisatisfiable if the following holds: “$\Gamma$ is satisfiable if and only if $\Delta$ is satisfiable”. For single-formula families $\Gamma = \{\varphi\}$ and $\Delta = \{\psi\}$, the following are equivalent conditions for equisatisfiability:

- There exist interpretations $I$ and $J$ such that $\varphi[I] = \psi[J]$.
- We have $\text{fun}_\varphi = \text{fun}_\psi$ for the induced functions.

A trivial example of equisatisfiable, but not equivalent formulas are $\varphi = P$ and $\psi = Q$ for $P \neq Q$.

**Theorem 814.** We give an explicit connection between propositional formulas and Boolean functions.

(a) The semantic equivalence $\equiv$ is an equivalence relation on the set $\text{Form}$ of all propositional formulas.

(b) The Lindenbaum-Tarski algebra $\text{Form} / \equiv$ of all propositional formulas with respect to semantic equivalence is bijective with the set of all Boolean functions of arbitrary arity. Both are provably Boolean algebras, but with very different proofs — the Lindenbaum-Tarski algebra is Boolean due to the purely syntactic proposition 919 and the set of all Boolean functions is a Boolean algebra due to the semantic proposition 866. This is another demonstration of theorem 900. See remark 920.
Proof.

**Proof of 814 (a).** Follows from the equivalences in definition 956.

**Proof of 814 (b).** Follows from the equivalences in definition 813 (e). □

**Proposition 815.** The following (and many more) are called Boolean equivalences because they are actually statements about our choice of standard Boolean operators. They are formulated here because the framework of propositional logic is more convenient for stating the equivalences. Note that most of these equivalences fail in intuitionistic logic.

For arbitrary propositional formulas $\varphi$ and $\psi$, the following semantic equivalences hold:

(a) **Negation** can be expressed via the falsum:

$$\neg \varphi \vDash \varphi \rightarrow \bot.$$  \hfill (260)

(b) **Negation** is an involution:

$$\neg \neg \varphi \vDash \varphi.$$  \hfill (261)

(c) **A material implication** is equivalent to its contrapositive:

$$\varphi \rightarrow \psi \vDash \neg \psi \rightarrow \neg \varphi.$$  \hfill (262)

(d) **A conditional** is a disjunction with the antecedent negated:

$$\varphi \rightarrow \psi \vDash \neg \varphi \lor \psi.$$  \hfill (263)

(e) **A biconditional** is a conjunction of conditionals:

$$\varphi \leftrightarrow \psi \vDash (\varphi \rightarrow \psi) \land (\psi \rightarrow \varphi).$$  \hfill (264)

(f) **The biconditional** is a conjunction of disjunctions:

$$\varphi \leftrightarrow \psi \vDash (\neg \varphi \lor \psi) \land (\neg \psi \lor \varphi).$$  \hfill (265)

(g) **The biconditional** is a disjunction of conjunctions:

$$\varphi \leftrightarrow \psi \vDash (\varphi \land \psi) \lor (\neg \varphi \land \neg \psi).$$  \hfill (266)

(h) **A biconditional** is equivalent its termwise negation:

$$\neg \varphi \leftrightarrow \neg \psi \vDash \varphi \leftrightarrow \psi.$$  \hfill (267)
(i) A negation of a biconditional is again a biconditional with one of the terms negated:

\[
\neg(\varphi \leftrightarrow \psi) \equiv \neg\varphi \leftrightarrow \psi \equiv \varphi \leftrightarrow \neg\psi.
\]  

\text{(268)}

Proof. The proofs follow directly from the table in \text{definition 800}. \hfill \Box

\textbf{Definition 816.} We sometimes want to substitute a propositional variable with another variable or even with a formula. This is akin to applying a Boolean function like \(x \lor y\) to different variables (e.g. to obtain \(x \lor x\)) or even concrete values (e.g. \(F \lor T\)), except that it is done on a purely syntactic level without involving any semantics involved.

It does not pose any technical difficulty to extend this definition beyond replacing a variable like it is usually done (e.g. \([\text{Aut20, def. 7.8}]\)). Not only that, we can then use this mechanism to define complicated rewriting rules as in \text{algorithm 821} (Perfect CNFs and DNFs) and have semantic equivalence automatically follow from \text{proposition 817}.

(a) We define the \textbf{substitution} of the propositional formula \(\theta\) with \(\chi\) in \(\varphi\) as

\[
\varphi[\theta \mapsto \chi] := \begin{cases} 
\chi, & \varphi = \theta \\
\varphi, & \varphi \neq \theta \text{ and } \varphi \in \{T, \bot\} \cup \text{Prop} \\
\neg\psi[\theta \mapsto \chi], & \varphi \neq \theta \text{ and } \varphi = \neg\psi \\
\psi_1[\theta \mapsto \chi] \circ \psi_2[\theta \mapsto \chi], & \varphi \neq \theta \text{ and } \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma.
\end{cases}
\]  

\text{(269)}

Note that it is not strictly necessary for \(\theta\) to be a subformula of \(\varphi\).

In the case where \(\theta\) is a single variable, if \(P \in \text{Var}(\varphi)\), then \(\varphi[P \mapsto \chi]\) is said to be an \textbf{instance} of \(\varphi\).

(b) We will now define \textbf{simultaneous substitution} of \(\theta_1, \ldots, \theta_n\) with \(\chi_1, \ldots, \chi_n\). We wish to avoid the case where \(\theta_k\) is a subformula of \(\chi_{k-1}\) and it accidentally gets replaced during \(\varphi[\theta_{k-1} \mapsto \chi_{k-1}][\theta_k \mapsto \chi_k]\).

Define

\[\text{Bound} := \text{Var}(\chi_1) \cup \ldots \cup \text{Var}(\chi_n).\]

and, for each variable \(P_k\) in \text{Bound}, pick a variable \(Q_k\) from \text{Prop} \setminus \text{Bound} (we implicitly assume the existence of enough variables in \text{Prop}). Let \(m\) be the \textbf{cardinality} of \text{Bound}. The simultaneous substitution can now be defined as

\[
\varphi[\theta_1 \mapsto \chi_1, \ldots, \theta_n \mapsto \chi_n] := \varphi[\theta_1 \mapsto \chi_1[P_1 \mapsto Q_1, \ldots, P_m \mapsto Q_m]] \circ \ldots \circ \varphi[\theta_n \mapsto \chi_n[P_1 \mapsto Q_1, \ldots, P_m \mapsto Q_m]]
\]

[270]

\textbf{Proposition 817.} If \(\theta\) is a subformula of \(\varphi\) and if \(\theta \models \chi\), then

\[
\varphi[\theta \mapsto \chi] \models \varphi.
\]  

By induction, this also holds for \textit{simultaneous substitution}.

402
Proof. We use structural induction on \( \varphi \): 

- If \( \varphi = \emptyset \), then \( \varphi[\emptyset \mapsto \chi] = \chi \) and, by definition, 
  \[ \varphi = \emptyset \models \chi = \varphi[\emptyset \mapsto \chi]. \]
- If \( \varphi \neq \emptyset \) and \( \varphi \in \{ \top, \bot \} \cup \text{Prop} \), then \( \varphi[\emptyset \mapsto \chi] = \varphi \) and (270) again holds trivially.
- If \( \varphi \neq \emptyset \) and \( \varphi = \neg \chi \) and if the inductive hypothesis holds for \( \chi \), then \( \varphi[\emptyset \mapsto \chi] = \neg \varphi[\emptyset \mapsto \chi] \). For any interpretation \( I \), 
  \[ (\varphi[\emptyset \mapsto \chi])[I] = \overline{(\psi[\emptyset \mapsto \chi])[I]} \overset{\text{ind.}}{=} \overline{\varphi[I]} = \varphi[I]. \]

Therefore, (270) holds in this case.
- If \( \varphi \neq \emptyset \) and \( \varphi = \psi_1 \circ \psi_2 \), \( \circ \in \Sigma \) and if the inductive hypothesis holds for both \( \psi_1 \) and \( \psi_2 \), then for any interpretation \( I \), 
  \[ (\varphi[\emptyset \mapsto \chi])[I] = (\psi_1[\emptyset \mapsto \chi])[I] \circ (\psi_2[\emptyset \mapsto \chi])[I] \overset{\text{ind.}}{=} \psi_1[I] \circ \psi_2[I] = \varphi[I]. \]

Therefore, (270) holds in this case also.

We have verified that (270) holds in all cases. \( \square \)

Remark 818. For semantical concepts, it is immaterial which element of an equivalence class we consider. Complete sets of Boolean operations allow us to represent each formula using a strict subset of the propositional constants, negation and connectives. Example 803 shows some concrete commonly used complete sets of Boolean operations. This is also the motivation for studying Lindenbaum-Tarski algebras.

This is useful in

- Reduction to normal forms such as the conjunctive normal form in algorithm 821 (Perfect CNFs and DNFs).
- Satisfiability proofs that rely on structural induction because it allows us to consider less cases in the induction.
- Having fewer rules in deductive systems. For example, we may choose to add (AX Pierce) to the axioms of the positive implicational derivation system and due to theorem 887 this derivation system would be able to emulate the classical derivation system.

Definition 819. We will now introduce conjunctive normal forms (CNF) and disjunctive normal forms (DNF) for propositional formulas. The concepts are related but distinct from that of lattice polynomials.
(a) The structure of these formulas is best described by the grammar schema:

\[
\begin{align*}
\langle \text{positive literal} \rangle &::= P \in \text{Prop} \\
\langle \text{negative literal} \rangle &::= \neg \langle \text{positive literal} \rangle \\
\langle \text{literal} \rangle &::= \langle \text{positive literal} \rangle \mid \langle \text{negative literal} \rangle \\
\langle \text{disjunct} \rangle &::= \langle \text{literal} \rangle \mid (\langle \text{literal} \rangle \lor \langle \text{disjunct} \rangle) \\
\langle \text{CNF} \rangle &::= \langle \text{disjunct} \rangle \land \langle \text{CNF} \rangle \\
\langle \text{conjunct} \rangle &::= \langle \text{literal} \rangle \mid (\langle \text{literal} \rangle \land \langle \text{conjunct} \rangle) \\
\langle \text{DNF} \rangle &::= \langle \text{conjunct} \rangle \lor \langle \text{DNF} \rangle
\end{align*}
\]

As usual, we utilize the convention in remark 808 and avoid excessive parentheses.

In this context, the terms **conjunct** and **disjunct** are commonly used to refer to sets of literals rather than the formulas containing them.

(b) Given a variable \( P \) and a Boolean value \( x \in \{T, F\} \), define

\[
P^x := \begin{cases} 
  P & x = T, \\
  \neg P & x = F.
\end{cases}
\]

(c) Given a finite sequence of distinct variables \( P_1, \ldots, P_n \), we say that a formula is in **perfect** CNF with respect to them if the following conditions hold:

(i) Every disjunct contains \( n \) literals and the \( k \)-th literal is either \( P_k \) or \( \neg P_k \).

(ii) The disjuncts are ordered lexicographically so that, for the \( k \)-th literals, \( L_k \leq R_k \) if either \( L_k \) is a negative literal or if both literals are positive.

A formula in perfect conjunctive normal form can be written as

\[
\prod_{(x_1, \ldots, x_n) \in B} P_{x_1}^1 \lor \cdots \lor P_{x_n}^n.
\]

Perfect DNFs are defined analogously. These additional conditions ensure uniqueness — see algorithm 821 (Perfect CNFs and DNFs).

**Example 820.** We list examples of formulas in conjunctive and disjunctive normal forms:

(a) The Boolean equivalence (263) allows us to convert the conditional \( P \rightarrow Q \) to \( \neg P \lor Q \), which is both in CNF and in DNF.

It is its own only disjunct, and it contains both variable, hence it is in perfect CNF.

The DNF is not perfect, however, because neither conditions definition 819 (c i) nor definition 819 (c ii) are satisfied.

(b) Consider instead the formula

\[
(\neg P \land \neg Q) \lor (\neg P \land Q) \lor (P \land Q).
\]

It is in perfect DNF, and it is equivalent to \( P \rightarrow Q \).
Algorithm 821 (Perfect CNFs and DNFs). Let \( f(x_1, ..., x_n) \) be an arbitrary Boolean function. We will build a formula in perfect disjunctive normal form and one in perfect conjunctive normal form. The induced function of both of these formulas will be \( f \). Both formulas are unique, as we will show.

(a) If \( f \) is constant, the constant itself is both a perfect CNF and DNF.

(b) Suppose that \( f \) is nonconstant and fix some propositional variables \( P_1, ..., P_n \). The following is a formula in perfect CNF whose induced function is \( f \):

\[
\bigwedge_{f(x_1, ..., x_n) = F} P_{x_1}^{x_1} \lor \cdots \lor P_{x_n}^{x_n}.
\]

(271)

Assuming that \( F < T \), we order the disjuncts with respect to the lexicographic order on the set \( \{T, F\}^n \) to which the tuples of Boolean values \( (x_1, ..., x_n) \) belong.

Dually, we construct the perfect DNF as

\[
\bigvee_{f(x_1, ..., x_n) = T} P_{x_1}^{x_1} \land \cdots \land P_{x_n}^{x_n},
\]

(272)

Proof of correctness. We will derive existence and uniqueness of perfect DNFs simultaneously from first principles. The derivation for CNFs is dual, but is more convoluted conceptually. Fix sequences of Boolean values \( a_1, ..., a_n \) and \( x_1, ..., x_n \). Then

\[
\begin{array}{c|c|c}
 x_k & a_k & P_k^{x_k}[a_k] \\
 F & F & T \\
 F & T & F \\
 T & F & F \\
 T & T & T \\
\end{array}
\]

Therefore, \( P_k^{x_k}[a_k] = T \) if and only if \( x_k = a_k \). Then

\[
\left(P_1^{x_1} \lor \cdots \lor P_n^{x_n}\right)[a_1, ..., a_n] = \begin{cases} T, & a_k = x_k \text{ for all } k = 1, ..., n, \\ F, & \text{otherwise}. \end{cases}
\]

Given some set \( B \subseteq \{T, F\}^n \), we have

\[
\left(\bigvee_{(x_1, ..., x_n) \in B} P_1^{x_1} \lor \cdots \lor P_n^{x_n}\right)[a_1, ..., a_n] = T
\]

if and only if there exists some tuple \( (x_1, ..., x_n) \in B \) such that \( x_k = a_k \) for every index \( k \). That is, if the tuple \( (a_1, ..., a_n) \) belongs to the complement \( B \). This leads us to the only possible definition

\[
B := \{(x_1, ..., x_n) \mid f(x_1, ..., x_n) = T\}.
\]

}\]
12.4. First-order logic

Definition 822. The idea of first-order predicate logic (we will omit “predicate” and only refer to “first-order logic”) is to create a formal language whose semantics (given by structures) support boolean operations and can quantify over all elements of an ambient universe. Unlike in propositional logic, there are different first-order logic languages.

The alphabet for a first-order logic language \( \mathcal{L} \) extends that of propositional logic and consists of two types of symbols (note that remark 804 holds here also).

Logical symbols

(a) The entirety of the propositional logic language except for the propositional variables.

(b) A nonempty at most countable alphabet of individual variables \( \text{Var} \), usually denoted by small Greek letters \( \xi_1, \xi_2, \ldots \) or \( \xi, \eta, \zeta \) — see remark 781 (a).

(c) The quantifiers \( Q = \{ \forall, \exists \} \):
   (i) The universal quantifier \( \forall \).
   (ii) The existential quantifier \( \exists \).
   (iii) The dot . for separating a quantifier from its formula.

The dot is not itself a quantifier and is not formally necessary — we use it only for readability.

(d) A symbol for formal equality \( \neq \). Equality is sometimes omitted by logicians, but examples of first-order languages without formal equality are obscure.

Non-logical symbols

(e) A set of functional symbols, \( \text{Fun} \), whose elements are usually denoted by \( f_1, f_2, \ldots \) or \( f, g \) or by symbols like \( \otimes \). In the latter case we usually use the infix notation discussed in remark 825 (a). Each functional symbol has an associated natural number called its arity, denoted by \( \# f \). Functional symbols with a zero arity are called constants.

(f) A set of predicate symbols, \( \text{Pred} \), whose elements are usually denoted by \( p_1, p_2, \ldots \) or by symbols like \( \leq \). Again, in the latter case we use infix notation. Predicate symbols also have an associated arity. Predicate symbols with zero arity are called propositional variables.

The logical symbols are common for all first-order languages. Thus, first-order languages differ by their non-logical symbols. The collection of functional and predicate symbols of a language are sometimes called its signature.

Definition 823. Similarly to the syntax of propositional logic, we define the syntax of a fixed first-order language \( \mathcal{L} \).
(a) Consider the following grammar schema:

\[
\begin{align*}
\langle \text{variable} \rangle & ::= \nu \in \text{Var} \\
\langle \text{connective} \rangle & ::= \circ \in \Sigma \\
\langle \text{quantifier} \rangle & ::= \forall \mid \exists \\
\langle \text{unary function} \rangle & ::= f \in \text{Fun}, \#f = 1 \\
\vdots \quad & \\
\langle n\text{-ary function} \rangle & ::= f \in \text{Fun}, \#f = n \text{ (standalone rule for each } n \text{)} \\
\vdots \quad & \\
\langle \text{unary predicate} \rangle & ::= p \in \text{Pred}, \#p = 1 \\
\vdots \quad & \\
\langle n\text{-ary predicate} \rangle & ::= p \in \text{Pred}, \#p = n \text{ (standalone rule for each } n \text{)} \\
\vdots \quad & \\
\langle \text{term} \rangle & ::= \langle \text{variable} \rangle \mid \\
& \quad \langle \text{unary function} \rangle \langle \langle \text{term} \rangle \rangle \mid \\
& \quad \vdots \quad \langle \text{unary predicate} \rangle \langle \langle \text{term} \rangle \rangle \mid \\
& \quad \vdots \quad \langle \text{n-ary function} \rangle \langle \langle \text{term}, \ldots, \langle \text{term} \rangle \rangle \rangle \mid \\
& \quad \vdots \quad \langle \text{atomic formula} \rangle ::= \top \mid \bot \mid \\
& \quad \langle \langle \text{term} \rangle \rangle \triangleq \langle \langle \text{term} \rangle \rangle \mid \\
& \quad \langle \text{unary predicate} \rangle \langle \langle \text{term} \rangle \rangle \mid \\
& \quad \vdots \quad \langle \text{n-ary predicate} \rangle \langle \langle \text{term}, \ldots, \langle \text{term} \rangle \rangle \rangle \mid \\
& \quad \vdots \quad \langle \text{formula} \rangle ::= \langle \text{atomic formula} \rangle \mid \\
& \quad \neg \langle \text{formula} \rangle \mid \\
& \quad \langle \langle \text{formula} \rangle \rangle \langle \text{connective} \rangle \langle \langle \text{formula} \rangle \rangle \mid \\
& \quad \langle \text{quantifier} \rangle \langle \text{variable} \rangle . \langle \text{formula} \rangle
\end{align*}
\]

In practice, we usually only have functions and predicates of specific arities. Note that we can have infinitely many functions, but only finitely many different arities. The theory of monoid actions is an example of a first-order language with infinitely many unary functional symbols, one constant and one binary functional symbol.

If we need the grammars to have a finite set of rules, except for having only finitely many different arities, we need to introduce appropriate naming conventions for variables, functions and predicates, analogously to the grammar schema of propositional logic.
We use the conventions in remark 808 regarding parentheses by extending them whenever appropriate.

In order to simplify notation, we also use the conventions in remark 825.

(b) The set Term of terms in \( \mathcal{L} \) is the language generated by this grammar schema with \( \langle \text{term} \rangle \) as a starting rule.

The grammar of first-order terms is unambiguous due to proposition 824, which makes it possible to perform proofs via theorem 792 (Structural induction on unambiguous grammars).

(c) If \( \tau \) and \( \chi \) are terms and \( \chi \) is a subword of \( \tau \), we say that \( \chi \) is a subterm of \( \tau \).

(d) For each term \( \tau \), we define its variables as

\[
\text{Var}(\tau) := \begin{cases} 
\{\xi\}, & \text{if } \tau = \xi \in \text{Var}, \\
\text{Var}(\tau_1) \cup \ldots \cup \text{Var}(\tau_n), & \text{if } \tau = f(\tau_1, \ldots, \tau_n).
\end{cases}
\]  

As in definition 806 (d), \( \text{Var} \) is ordered by the position of the first occurrence of a variable.

(e) A term \( \tau \) is called a ground term if \( \text{Var}(\tau) = \emptyset \). Ground terms are also called closed terms.

(f) The set Atom of atomic formulas in \( \mathcal{L} \) is the language generated by this grammar schema with \( \langle \text{atomic formula} \rangle \) as a starting rule.

(g) The set Form of formulas in \( \mathcal{L} \) is the language generated by this grammar schema with \( \langle \text{formula} \rangle \) as a starting rule.

The atomic formulas are the ones generated from \( \langle \text{atomic formula} \rangle \).

The grammar of first-order formulas is unambiguous as shown by proposition 824.

See example 838 for examples of first-order formulas.

(h) If \( \varphi \) and \( \psi \) are formulas and \( \psi \) is a subword of \( \varphi \), we say that \( \psi \) is a subformula of \( \varphi \).

(i) If \( \varphi \) is a formula, if \( \tau \) is a term and if \( \tau \) is a subword of \( \varphi \), we say that \( \tau \) is a term of \( \varphi \).

(j) The free variables of a formula are defined as

\[
\text{Free}(\varphi) := \begin{cases} 
\emptyset, & \varphi \in \{\top, \bot\} \\
\text{Var}(\tau_1) \cup \ldots \cup \text{Var}(\tau_n), & \varphi = p(\tau_1, \ldots, \tau_n) \\
\text{Var}(\tau_1) \cup \text{Var}(\tau_2), & \varphi = \tau_1 \equiv \tau_2, \\
\text{Free}(\psi), & \varphi = \neg \psi, \\
\text{Free}(\psi_1) \cup \text{Free}(\psi_2), & \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma, \\
\text{Free}(\psi) \setminus \{\xi\}, & \varphi = Q \xi \cdot \psi, Q \in \{\forall, \exists\}
\end{cases}
\]  

408
(k) A formula $\varphi$ is called a **ground formula** if $\text{Free}(\varphi) = \emptyset$. Ground formulas are also called **closed formulas** or **sentences** (unlike in propositional logic where all formulas are called sentences — see definition 806 (b)).

We will not restrict our attention only to closed formulas, and we will even rely on implicit quantification as mentioned in remark 781 (g). That being said, certain important theorems like theorem 846 (First-order semantic deduction theorem) and theorem 882 (Syntactic deduction theorem) require some of the formulas to be closed, so we will often follow this restriction.

(l) Dually, the **bound variables** of a formula are defined as

$$
\text{Bound}(\varphi) :=
\begin{cases}
\emptyset, & \varphi \text{ is atomic,} \\
\text{Bound}(\psi), & \varphi = \neg \psi, \\
\text{Bound}(\psi_1) \cup \text{Bound}(\psi_2), & \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma, \\
\text{Bound}(\psi) \cup \{\xi\}, & \varphi = Q \xi, Q \in \{\forall, \exists\}.
\end{cases}
$$

(m) Finally, the set of all variables of a formula $\varphi$ is

$$
\text{Var}(\varphi) := \text{Free}(\varphi) \cup \text{Bound}(\varphi).
$$

**Proposition 824.** *The grammars of first-order terms and of first-order formulas are unambiguous.*

**Proof.** The proof is more complicated, but similar in spirit to proposition 807.

**Remark 825.** In order to simplify exposition, we use the following conventions

(a) Binary functional symbols are often written using **infix notation**, i.e.

$$
\zeta \doteq \xi + \eta
$$

rather than the **prefix notation**

$$
\zeta \doteq + (\xi, \eta).
$$

This also applies to predicates — we write $\xi \sim \eta$ rather than $\sim (\xi, \eta)$.

(b) Negation of an infix binary predicate symbol $\sim$ can be written more simply as

$$
\xi \sim \eta
$$

rather than

$$
\neg (\xi \sim \eta).
$$

409
(c) If \( \sim \) is a binary predicate, to further shorten notation, we write
\[
\forall \xi \sim \eta. \varphi
\]
as a shorthand for
\[
\forall \xi. (\xi \sim \eta \rightarrow \varphi)
\]
and
\[
\exists \xi \sim \eta. \varphi
\]
as a shorthand for
\[
\exists \xi. (\xi \sim \eta \land \varphi).
\]
This is called relativization of the quantifier and is immensely useful when working with heterogeneous objects or even in set theory.

(d) We sometimes want to specify not only existence, but also uniqueness. This is the case in (350), for example. It is conventional to write
\[
\exists! \xi. \varphi
\]
as a shorthand for
\[
\exists \xi. (\varphi \land (\forall \eta. \varphi[\xi \mapsto \eta] \rightarrow (\xi \equiv \eta))).
\]

(e) We only add to the language itself the functional and predicate symbols that are necessary for our desired axioms — see definition 905. We can define additional functions and predicates in terms of these, but we avoid using them as much as possible when writing formulas in the object language. For example, we avoid adding the functional symbols \( \text{ord}(A) \) and \( \text{card}(A) \) or even \( \cup \) and \( \cap \) to ZFC.

If needed, we can consider these new functions and predicates to be abbreviations for more verbose terms and formulas as described in remark 841.

As for remark 808, both of these conventions exist only in the metalanguage and the formulas themselves are assumed to have the former form within the object language.

[Aut20] def. 16.1 **Definition 826.** Fix a first-order logic language \( \mathcal{L} \). A **structure** for \( \mathcal{L} \) is a pair \( \mathcal{X} = (X, I) \), where

(a) \( X \) is a nonempty set called the **domain** or **universe** of the structure \( \mathcal{X} \). See remark 827.

(b) The **interpretation** \( I \) of the structure \( \mathcal{X} \) is a **function** that is defined on the signature of \( \mathcal{L} \) and satisfies the following conditions:

(i) For every \( n \)-ary function symbol \( f \), its interpretation is a function with signature \( I(f) : X^n \rightarrow X \).
(ii) For every \(n\)-ary predicate \(p\), its interpretation is a \(n\)-ary Boolean-valued function with signature \(I(p) : X^n \to \{T, F\}\). A tuple \((x_1, \ldots, x_n)\) satisfies \(p\) if \(p(x_1, \ldots, x_n) = T\).

It is conventional to define the interpretation of a predicate to be a relation \(I(p) \subseteq X^n\) (see e.g.), however it is more convenient for us to work with Boolean-valued functions. The two approaches are equivalent as explained in remark 797.

Unlike in the rest of this document, when dealing with first-order structures, it is important to distinguish between the structure \(\mathcal{X}\) as a pair and its domain \(X\) as a set. See remark 834.

Remark 827. If we allow for the domain of a structure to be empty, we would have to reformulate a lot of important theorems (e.g. see the proof of proposition 839 (b)), which would complicate compatibility between semantics and deductive systems.

See proposition 855 (d) for a context where empty sets are justified as domains of first-order structures.

Definition 828. Fix a structure \(\mathcal{X} = (X, I)\) for a first-order logic language \(\mathcal{L}\).

(a) A variable assignment for the variables of \(\mathcal{L}\) is any function \(v : \text{Var} \to X\) (loosely similar to propositional interpretations).

(b) For every variable \(\xi\) and every domain element \(x \in X\) we also define the modified assignment at \(\xi\) with \(x:\)

\[
u_{\xi \mapsto x}(\xi) := \begin{cases} x, & \xi = \xi, \\ v(\xi), & \xi \neq \xi. \end{cases}
\]

We can also modify the value at \(\xi\) with another variable, e.g.

\[
u_{\xi \mapsto \eta}(\xi) := \begin{cases} v(\eta), & \xi = \xi, \\ v(\xi), & \xi \neq \xi. \end{cases}
\]

Inductively,

\[
u_{\xi_1 \mapsto x_1, \ldots, \xi_n \mapsto x_n}(\eta) := \nu((\ldots (\nu_{\xi_1 \mapsto x_1}) \ldots)_{\xi_n \mapsto x_n})(\eta).
\]

Except for semantics of quantification, these are also used in other places like proposition 839 and remark 829.

(c) The valuation of a term \(\tau\) is a value in the domain \(X\) given by

\[
\tau[v] := \begin{cases} v(\xi), & \tau = \xi \in \text{Var}, \\ I(f)(\tau_1[v], \ldots, \tau_n[v]), & \tau = f(\tau_1, \ldots, \tau_n). \end{cases}
\]
We extend the classical propositional valuations from definition 810. The (classical) valuation of a formula \( \varphi \) is a Boolean value given by

\[
\varphi[\psi] := \begin{cases} 
T, & \varphi = T, \\
F, & \varphi = \bot, \\
\tau_1[\psi] = \tau_2[\psi], & \varphi = \tau_1 \equiv \tau_2, \\
I(p)(\tau_1[\psi], ..., \tau_n[\psi]), & \varphi = p(\tau_1, ..., \tau_n), \\
\psi[\psi], & \varphi = \neg\psi, \\
\psi_1[\psi] \circ \psi_2[\psi], & \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma, \\
\vee\{\psi[u_{\xi \mapsto x}] \mid x \in X\}, & \varphi = \forall \xi . \psi, \\
\wedge\{\psi[u_{\xi \mapsto x}] \mid x \in X\}, & \varphi = \exists \xi . \psi.
\end{cases}
\] (278)

The rules for evaluating constants, negations and connectives are a direct extension of the rules for propositional logic.

It is important that the domain is nonempty because \( \wedge \emptyset = T \), which directly contradicts our intent of defining \( \exists \) as a quantifier for existence.

**Remark 829.** Somewhat similar to remark 811, if we know that \( \text{Free}(\varphi) \subseteq \{\xi_1, ..., \xi_n\} \), we know that the valuation \( \varphi[\psi] \) only depends on the values \( \psi(\xi_1), ..., \psi(\xi_n) \). This allows us to introduce the shorthand

\[
\varphi[\xi_1 \mapsto x_1, ..., \xi_n \mapsto x_n]
\] (279)

or even

\[
\varphi[x_1, ..., x_n]
\] (280)

for

\[
\varphi[u_{\xi_1 \mapsto x_1, ..., \xi_n \mapsto x_n}]
\]

because the variable assignment \( \psi \) plays no role here.

When using either of these shorthand notations, we implicitly assume that \( \text{Free}(\varphi) \subseteq \{\xi_1, ..., \xi_n\} \).

When \( \varphi = p(\xi_1, ..., \xi_n) \) is a predicate formula, the shorter notation (280) translates to

\[
p[\alpha_1, ..., \alpha_n] = I(p)(\alpha_1, ..., \alpha_n) \in \{T, F\}
\]

and analogously for function terms we have

\[
f[\alpha_1, ..., \alpha_n] = I(f)(\alpha_1, ..., \alpha_n) \in X.
\]

Of course, we avoid this notation for formulas like \( p(f(\xi)) \) because \( p[\xi] \) would mean \( I(p)(I(f)(\xi)) \) rather than \( I(p)(\alpha) \), which would be confusing.

We apply this notation for terms and, in particular, functions.

We also sometimes use the shortened notation

\[
\varphi[\tau_1, ..., \tau_n]
\] (281)

for substituting terms in formulas.
Definition 830. A first-order equation is a formula of the form

\[ f(\xi_1, \ldots, \xi_n) \equiv g(\xi_1, \ldots, \xi_n), \tag{282} \]

where both \( f(\xi_1, \ldots, \xi_n) \) and \( g(\xi_1, \ldots, \xi_n) \) are functional symbols with the same free variables. Given a structure \( \mathcal{X} = (X, I) \), we call the elements of the set defined by this formula solutions. That is, we say that the tuple \((x_1, \ldots, x_n)\) is a solution to the equation (282) if

\[ f[x_1, \ldots, x_n] = g[x_1, \ldots, x_n]. \]

We can actually replace \( f \) and \( g \) with more general terms \( \tau \) and \( \sigma \), in which case the equation becomes \((\tau \equiv \sigma)\) and is satisfied if

\[ \tau[x_1, \ldots, x_n] = \sigma[x_1, \ldots, x_n]. \]

Example 831. A remarkable portion of mathematics concerns the study of different types of equations (even though they are not generally restricted to equations in first-order logic). The reason for this is that equations provide a simple way to specify rich semantic structure using simple syntactic objects.

- Matrix theory can be regarded as the study of linear equations. See ?? ([UNDEFINED]).
- Differential equations is aptly named since it studies equations in functional spaces concerning functions and their derivatives. See ?? ([UNDEFINED]).
- Roots of generalized derivatives are studied in optimization. See section 2.9 (Non-smooth derivatives).
- Diophantine equations are studied in number theory. See section 1.2 (Integers).
- Fixed points of functions are studied in different branches of mathematics. See theorem 390 (Banach’s fixed point theorem) or theorem 1259 (Knaster-Tarski theorem).
- Affine varieties, which are sets of roots of polynomials, are studied in algebraic geometry. See ?? ([UNDEFINED]).

Definition 832. Fix a first-order logic language \( \mathcal{L} \). We introduce notions analogous to propositional semantics:

(a) Given a structure \( \mathcal{X} = (X, I) \), a variable assignment \( \nu \) and a set \( \Gamma \) of first-order formulas, we say that the variable assignment \( \nu \) satisfies \( \Gamma \) and we write \( \mathcal{X} \models_{\nu} \Gamma \) if, for every formula \( \gamma \in \Gamma \) we have \( \gamma[\nu] = T \).

If every variable assignment in \( \mathcal{X} \) satisfies \( \Gamma \), we say that \( \mathcal{X} \) itself satisfies \( \Gamma \) or that \( \mathcal{X} \) is a model of \( \Gamma \) and write \( \mathcal{X} \models \Gamma \) (or simply \( \mathcal{X} \models \Gamma \) if the interpretation is clear from the context).

Analogously to definition 813 (a), we say that \( \Gamma \) is satisfiable if there exists a model for \( \Gamma \).
(b) We say that the set of formulas $\Gamma$ **entails** the set of formulas $\Delta$ and write $\Gamma \vdash \Delta$ if every model of $\Gamma$ is also a model of $\Delta$.

(c) The formula $\varphi$ is a **tautology** if every structure is a model of $\varphi$.

(d) Dually, $\varphi$ is a **contradiction** is no structure is a model of $\varphi$.

(e) As in the simplest case with **propositional semantical equivalence**, we say that $\Gamma$ and $\Delta$ are **semantically equivalent** and write $\Gamma \equiv \Delta$ if both $\Gamma \vdash \Delta$ and $\Delta \vdash \Gamma$.

(f) Again as in the simplest case with **propositional equisatisfiability**, we say that the sets of formulas $\Gamma$ and $\Delta$ are **equisatisfiable** when it holds that $\Gamma$ is satisfiable if and only if $\Delta$ is satisfiable.

**Proposition 844** (b) provides an important example of equisatisfiable formulas that are not equivalent.

**Remark 833.** It is now clear that the **propositional logic language** can be regarded as a degenerate first-order logic language with at most countably many predicate symbols, all of arity 0, and no functional symbols. Thus, first-order logic is indeed an extension of propositional logic.

**Remark 834.** In first-order logic, **models** are defined as pairs $\mathcal{M} = (X, I)$. Each area of mathematics has its own conventions and models are usually specified as simply as possible without being unambiguous (and sometimes even beyond non-ambiguity).

A popular convention is to use compatible letters like we did with $X$ and $\mathcal{X}$ or $G$ and $\mathcal{G}$, where the structure itself is named using calligraphic letters while the domain is named using the corresponding capital letter in normal font. This only works very simple cases where we can say “Let $\mathcal{P} = (P, \leq)$ be a partially ordered set”.

The language of the **theory of groups** has a signature consisting of three functional symbols and no predicate symbols. Specifying a structure for this language is thus the same as specifying a quadruple $\mathcal{G} = (G, e, (\cdot)^{-1}, \cdot)$. We usually specify only the domain $G$ and the basic structure needed to avoid ambiguity, for example “Let $(G, \cdot)$ be a group”. This is technically wrong, but it is both convenient and conventional. The rest of the definition of the group can easily be inferred. In case of ambiguity, the simplest disambiguation is to use lower indices with the name of the structure, e.g. $+_G$ and $+_H$ may be the addition operation in different abelian groups.

Furthermore, stating that $(G, \cdot, \leq, \mathcal{T})$ is a totally ordered topological group is cumbersome and can even raise questions; for example, is $\mathcal{T}$ the **order topology** or just an arbitrary **group topology**?
12.5. First-order satisfiability

**Definition 835.** As in propositional logic, we sometimes want to perform substitution, however we have different types of syntactic objects (terms and formulas) which have different substitution rules. The notion of free and bound variables further complicates us — see for example the problems outlined in remark 836. In particular, this means that an analogous to proposition 817 theorem cannot longer justify substitution as it is done in [UNDEFINED] — we can have weaker statements as in proposition 817 that implicitly rely on variable renaming in order to hold. This implies that it is of no practical use to define substitution of a first-order subformula inside another formula as it is done in definition 816. Instead, we concert ourselves with substituting variables — propositional variables with first-order formulas and first-order variables with first-order terms. Furthermore, since this does not complicate us, we allow substituting arbitrary terms rather than only first-order variables.

While substituting a propositional variable is the syntactic analog to applying Boolean functions to different variables or propositional constants, substituting a first-order variable can express applying arbitrary functions to different first-order variables or arbitrary constants. For example, in a suitable language, we can apply \( \log(x) \) to the constant \( c \) by substituting \( x \) with \( c \) to obtain the ground term \( \log(c) \).

As in definition 816, we define different kinds of (single) substitution in more generality that in e.g. [Aut20, def. 15.25]. Where applicable, simultaneous substitution is defined via the same trick as in definition 816.

(a) Let \( \varphi \) be a propositional formula with variables \( \text{Var}(\varphi) = \{P_1, \ldots, P_n\} \). For brevity, denote \( V := \text{Var}(\varphi) \). Let \( \Theta = \{\delta_1, \ldots, \delta_n\} \) be a set of first-order formulas.

It does not make sense to replace a single propositional variable by a single formula. Furthermore, a first-order formula \( \delta_k \) cannot possibly contain any of the propositional variables \( P_1, \ldots, P_n \). This allows us to introduce a simplification of the simultaneous substitution based on (269) as

\[
\phi[V \mapsto \Theta] := \begin{cases} 
\varphi, & \varphi \in \{\top, \bot\} \\
\delta_k, & \varphi = \delta_k \text{ for some } k = 1, \ldots, n \\
\lnot\psi[V \mapsto \Theta], & \varphi = \lnot\psi \\
\psi_1[V \mapsto \Theta] \circ \psi_2[V \mapsto \Theta], & \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma.
\end{cases}
\] (283)

As in definition 816, it is not strictly necessary for any of the variables to belong to \( \text{Var}(\varphi) \).

(b) We define the substitution of the first-order term \( \kappa \) with \( \mu \) in the term \( \tau \) as

\[
\tau[\kappa \mapsto \mu] := \begin{cases} 
\mu, & \tau = \kappa, \\
\tau, & \tau \neq \kappa \text{ and } \tau \in \text{Var}, \\
f(\tau_1[\kappa \mapsto \mu], \ldots, \tau_n[\kappa \mapsto \mu]), & \tau \neq \kappa \text{ and } \tau = f(\tau_1, \ldots, \tau_n).
\end{cases}
\] (284)

It is not strictly necessary for \( \kappa \) to be a subterm of \( \tau \).
(c) This case is more complicated. We define the substitution of the term \( \kappa \) with the term \( \mu \) in the first-order formula \( \varphi \) as

\[
\varphi[\kappa \mapsto \mu] := \begin{cases} 
\varphi, & \varphi \in \{\top, \bot\}, \\
p(\tau_1[\kappa \mapsto \mu], ..., \tau_n[\kappa \mapsto \mu]), & \varphi = p(\tau_1, ..., \tau_n), \\
\tau_1[\kappa \mapsto \mu] \equiv \tau_2[\kappa \mapsto \mu], & \varphi = \tau_1 \equiv \tau_2, \\
\neg \psi[\kappa \mapsto \mu], & \varphi = \neg \psi, \\
\psi_1[\kappa \mapsto \mu] \circ \psi_2[\xi \mapsto \mu], & \varphi = \psi_1 \circ \psi_2, \quad o \in \Sigma, \\
(\dagger), & \varphi = Q \xi \cdot \psi, Q \in \{\forall, \exists\}, 
\end{cases}
\]  

(285)

where

\[
(\dagger) := \begin{cases} 
\varphi, & \xi \in \text{Var}(\kappa), \\
Q \xi \cdot (\psi[\kappa \mapsto \mu]), & \xi \notin \text{Var}(\kappa) \cup \text{Var}(\mu), \\
Q \eta \cdot (\psi[\xi \mapsto \eta][\kappa \mapsto \mu]), & \xi \notin \text{Var}(\kappa) \text{ and } \xi \in \text{Var}(\mu) \text{ and } \eta \notin \text{Var}(\kappa) \cup \text{Var}(\mu) \cup \text{Var}(\psi).
\end{cases}
\]  

(286)

(287)

(288)

In (288), we chose a new variable \( \eta \). We implicitly assume that there exist enough variables in the language, so that we can find \( \eta \) that satisfies the condition. In order to fully avoid nondeterminism in the choice of \( \eta \), we can pick a well-ordering on the set \( \text{Var} \) and always choose \( \eta \) to be the smallest variable not present in \( \varphi \). This rule is called **renaming of the bound variables** \( \xi \) to \( \eta \) and is done to mitigate capturing as described in remark 836 (a).

We could avoid the rule for renaming (as it is done in [Aut20, def. 15.25]), however renaming both free and bound variables is natural and is often done in practice. For example, consider the Peano arithmetic formula “there exists \( n \) such that \( nm \) is even". Note that the bound variable \( n \) is renamed to \( k \) and the free variable \( m \) to \( n \) in the larger formula “for every \( n \) there exists \( k \) such that \( kn \) is even".

The rule (286) may seem redundant, but when doing inductive proofs (e.g. the proof of proposition 839), we usually need to separately consider the cases where \( \xi \in \text{Var}(\kappa) \) and \( \xi \notin \text{Var}(\kappa) \setminus \text{Var}(\mu) \) and the rule (287) being trivial simplifies the proofs.

See remark 837 regarding the additional parentheses in (288).

See example 838 for examples of applying the different quantifier rules.

Remark 836. The renaming rule (288) is designed to mitigate the following two problems (compared to (287)):

(a) Renaming mitigates “capturing” free variables as in

\[
(\forall \eta \cdot p(\xi, \eta))[\xi \mapsto \eta] = \forall \eta \cdot p(\eta, \eta)
\]

by instead producing, up to a choice of new variables, the formula

\[
(\forall \eta \cdot p(\xi, \eta))[\xi \mapsto \eta] = \forall \zeta \cdot p(\eta, \zeta).
\]

416
(b) Renaming mitigates “colliding” multiple bound variables as in
\[
(\forall \xi . \forall \eta . p(\xi, \eta))\[\xi \mapsto \eta\] = \forall \xi . \forall \eta . p(\eta, \eta)
\]
by instead producing, up to a choice of new variables, the formula
\[
(\forall \xi . \forall \eta . p(\xi, \eta))\[\xi \mapsto \eta\] = \forall \zeta . \forall \sigma . p(\zeta, \sigma).
\]

Remark 837. When performing substitution, it is sometimes convenient to add additional parentheses to avoid ambiguity. For example, while parentheses around quantifier expressions are not necessary by the syntax of first-order logic, adding such parentheses helps avoid the ambiguity in
\[
\forall \xi . p(\xi, \eta)(\eta \mapsto \zeta).
\]
Instead, we either write
\[
(\forall \xi . p(\xi, \eta))[\eta \mapsto \zeta]
\]
or
\[
\forall \xi . (p(\xi, \eta)(\eta \mapsto \zeta)).
\]
This convention is only part of the metasyntax and the parentheses are not part of the syntax of the formulas themselves.

Example 838. The following term substitutions should justify the distinct cases in (285):

(a) The trivial case without actual substitution:
\[
(\forall \xi . p(\xi, \eta))[\xi \mapsto \eta] = (\forall \xi . p(\xi, \eta))[\xi \mapsto \eta] = \forall \xi . p(\xi, \eta).
\]

Example 838 (e) demonstrates that this does not work for nested substitution.

(b) A simple substitution without renaming:
\[
(\forall \xi . p(\xi, \eta))[\eta \mapsto \zeta] = \forall \xi . (p(\xi, \eta)(\eta \mapsto \zeta)) = \forall \xi . p(\xi, \zeta).
\]

(c) A simple renaming without actual substitution:
\[
(\forall \xi . p(\xi, \eta))[\eta \mapsto \xi] = \forall \xi . (p(\xi, \eta)(\xi \mapsto \xi)) = \forall \zeta . p(\zeta, \xi).
\]
(d) Example 838 (c), but with $\mu$ in (285) containing $\xi$ indirectly:

$$\left( \forall \xi \cdot p(\xi, \eta) \right)[\eta \mapsto f(\xi)] \overset{288}{=} \forall \xi \cdot \left( p(\xi, \eta)[\xi \mapsto \xi][\eta \mapsto f(\xi)] \right) \overset{285}{=} \forall \xi \cdot p(\xi, f(\xi)).$$

(e) Only renaming with multiple quantifiers which shows the limitations of (286):

$$\left( \forall \eta \cdot \exists \xi \cdot p(\xi, \eta) \right)[\xi \mapsto \eta] \overset{288}{=} \forall \xi \cdot \left( \left( \forall \eta \cdot p(\xi, \eta) \right)[\eta \mapsto \xi][\xi \mapsto \eta] \right) \overset{285}{=} \forall \xi \cdot \left( \forall \eta \cdot p(\xi, \xi) \right) \overset{286}{=} \forall \xi \cdot \forall \eta \cdot p(\xi, \xi).$$

(f) Both renaming and substitution with multiple quantifiers:

$$\left( \forall \eta \cdot \left( p(\xi, \eta) \lor \forall \xi \cdot p(\xi, \eta) \right) \right)[\xi \mapsto \eta] \overset{288}{=} \forall \xi \cdot \left( \left( \forall \eta \cdot p(\xi, \eta) \lor \exists \xi \cdot p(\xi, \eta) \right) \right)[\eta \mapsto \xi][\xi \mapsto \eta] \overset{285}{=} \forall \xi \cdot \left( \forall \eta \cdot p(\xi, \xi) \lor \exists \xi \cdot p(\xi, \xi) \right) \overset{286}{=} \forall \xi \cdot \forall \eta \cdot \left( \exists \xi \cdot p(\xi, \xi) \right).$$

(g) Substitution of more general terms than variables with renaming of term’s variables:

$$\left( \forall \xi \cdot p(\xi, \eta, f(\eta)) \right)[f(\eta) \mapsto g(\eta, \xi)] \overset{288}{=} \forall \xi \cdot \left( p(\xi, \eta, f(\eta)) \right)[\xi \mapsto \xi][f(\eta) \mapsto g(\eta, \xi)] \overset{285}{=} \forall \xi \cdot \forall \eta \cdot \left( p(\xi, \eta, g(\eta, \xi)) \right).$$

Proposition 839. We will show how syntactic renaming is compatible with a certain “semantic renaming”.

Fix a first-order language $\mathcal{L}$, a structure $\mathcal{X} = (X, I)$ on $\mathcal{L}$ and a variable assignment $\nu$ in $\mathcal{X}$.

(a) For any term $\tau$ and any two variables $\xi$ and $\eta$, we have

$$\tau \nu_{\xi \mapsto \eta} = \left( \tau[\xi \mapsto \eta] \right) \nu \|.$$

(b) For any formula $\varphi$, any variable $\xi$ and any other variable $\eta$ not in $\mathbf{Var}(\varphi)$ we have

$$\varphi \nu_{\xi \mapsto \eta} = \left( \varphi[\xi \mapsto \eta] \right) \nu \|.$$
Proof. In both cases, we use structural induction on the definition of the substitution.

**Proof of 839 (a).**

- If \( \tau = \xi \), then
  \[
  \tau[\xi \mapsto \eta] = \xi[\xi \mapsto \eta] = \eta
  \]
  and (289) follows directly.
- If \( \tau \) is a variable and \( \tau \neq \xi \), then
  \[
  \tau[\xi \mapsto \eta] = \tau
  \]
  and (289) again holds trivially.
- If \( \tau = f(\tau_1, \ldots, \tau_n) \) and if the inductive hypothesis holds for \( \tau_1, \ldots, \tau_n \), then
  \[
  \left( \tau[\xi \mapsto \eta] \right)[\nu] = \left( f(\tau_1[\xi \mapsto \eta], \ldots, \tau_n[\xi \mapsto \eta]) \right)[\nu] = \]
  \[
  = I(f)\left(\left( \tau_1[\xi \mapsto \eta] \right)[\nu], \ldots, \left( \tau_n[\xi \mapsto \eta] \right)[\nu] \right) \]
  \[
  = I(f)\left( \tau_1[\nu_{\xi \mapsto \eta}], \ldots, \tau_n[\nu_{\xi \mapsto \eta}] \right) = \]
  \[
  = \left( f(\tau_1[\nu_{\xi \mapsto \eta}], \ldots, \tau_n[\nu_{\xi \mapsto \eta}] \right) = \]
  \[
  = \tau[\nu_{\xi \mapsto \eta}].
  \]

In all cases, (289) holds.

**Proof of 839 (b).**

- If \( \phi \in \{ \top, \bot \} \), then \( \phi \) has no subterms and thus (290) holds vacuously.
- If \( \phi = p(\tau_1, \ldots, \tau_n) \), then since (289) holds for all \( \tau_k \), we have
  \[
  \left( \tau_k[\xi \mapsto \eta] \right)[\nu] = \tau_k[\nu_{\xi \mapsto \eta}] \]
  and thus
  \[
  I(p)\left( \left( \tau_1[\xi \mapsto \eta] \right)[\nu], \ldots, \left( \tau_n[\xi \mapsto \eta] \right)[\nu] \right) \]
  \[
  = I(p)\left( \tau_1[\nu_{\xi \mapsto \eta}], \ldots, \tau_n[\nu_{\xi \mapsto \eta}] \right). \]

Therefore,
\[
\left( \phi[\xi \mapsto \eta] \right)[\nu] = \left( p(\tau_1[\xi \mapsto \eta], \ldots, \tau_n[\xi \mapsto \eta]) \right)[\nu] = \]
\[
= \left( p(\tau_1, \ldots, \tau_n) \right)[\nu_{\xi \mapsto \eta}] = \]
\[
= \phi[\nu_{\xi \mapsto \eta}].
\]

- The case \( \phi = \tau_1 \neq \tau_2 \) is proved analogously.
• The cases $\varphi = \neg \psi$ and $\varphi = \psi_1 \circ \psi_2$ are proved in a straightforward manner.

• Let $\varphi = \forall \zeta . \psi$, where the inductive hypothesis holds for $\psi$. We consider three cases
  
  – Suppose that $\zeta = \xi$. By definition, we have
    \[ \varphi[\xi \mapsto \eta] = \varphi, \]
    hence (290) holds trivially.

  – Suppose that $\zeta \neq \xi$. It follows that
    \[ \varphi[\xi \mapsto \eta] = \forall \zeta . \left( \psi[\xi \mapsto \eta] \right). \]

    * If $\left( \varphi[\xi \mapsto \eta] \right)[v] = T$, by definition of quantifier formula valuation, for any $x \in X$ we have
      \[ \left( \forall \zeta . \left( \psi[\xi \mapsto \eta] \right) \right)[v] = \left( \psi[\xi \mapsto \eta] \right)[v_{\xi \mapsto x}] = T. \tag{291} \]
      On the other hand, by the inductive hypothesis,
      \[ \left( \psi[\xi \mapsto \eta] \right)[v] = \psi[u_{\xi \mapsto \eta}], \]
      and, as a special case, for any $x \in X$,
      \[ \left( \psi[\xi \mapsto \eta] \right)[v_{\xi \mapsto \eta \mapsto x}] = \psi[u_{\xi \mapsto \eta \mapsto x}]. \tag{292} \]
      Combining (291) and (292), we obtain
      \[ \left( \varphi[\xi \mapsto \eta] \right)[v] = \left( \psi[\xi \mapsto \eta] \right)[v_{\xi \mapsto x}] = \psi[u_{\xi \mapsto \eta \mapsto x}] = \varphi[u_{\xi \mapsto \eta}], \quad \text{for all } x \in X, \]
    which proves the case.

    * If $\left( \varphi[\xi \mapsto \eta] \right)[v] = F$, then there exists $x \in X$ such that
      \[ \left( \psi[\xi \mapsto \eta] \right)[v_{\xi \mapsto x}] = F. \]
      Since (292) holds by the inductive hypothesis, we have
      \[ \psi[u_{\xi \mapsto \eta \mapsto x}] = F \]
      for the same $x$.
      It follows that $\varphi[u_{\xi \mapsto \eta}] = F$, which proves the case.
We can prove the case $\varphi = \exists \zeta . \psi$ using double negation on the previous case.

In all cases, (290) holds. $\square$

**Proposition 840.** Analogously to proposition 817, we will show that all substitutions defined in definition 835 types of substitution preserve the corresponding semantics.

By induction, this proposition also holds for simultaneous substitution.

Fix a structure $\mathcal{X} = (X, I)$ and a variable assignment $v$.

(a) As in definition 835 (a), let $\varphi$ be a propositional formula with variables $V = \{P_1, \ldots, P_n\}$ and let $\Theta = \{\theta_1, \ldots, \theta_n\}$ be a set of first-order formulas.

Furthermore, let $J$ be a propositional interpretation such that, for all $k = 1, \ldots, n$,

$$P_k[J] = \theta_k[v].$$

Then

$$\left( \varphi[V \mapsto \Theta] \right)[v] = \varphi[J].$$

In particular, $\models \varphi$ in the sense of definition 813 (c) implies $\models \varphi[V \mapsto \Theta]$ in the sense of definition 832 (c).

(b) Let $\tau$ be a first-order term and let $x$ be a subterm of $\tau$. Let $\mu$ be another term such that

$$\mu[v] = x[v].$$

Then

$$\tau[x \mapsto \mu][v] = \tau[v].$$

(c) Let $\varphi$ be a first-order formula and let $x$ be a term of $\varphi$. Let $\mu$ be another term such that

$$\mu[v] = x[v].$$

Then

$$\varphi[x \mapsto \mu][v] = \varphi[v].$$

**Proof.** In all cases, we use structural induction by the definition of the substitution. The inductive hypothesis for a formula is that the proposition holds for arbitrary substitutions and valuations.

**Proof of 840 (a).** Let $\varphi$ be a propositional formula.

- If $\varphi \in \{\top, \bot\}$, no substitution is performed and thus (294) holds trivially.
- If $\varphi = P_k$ for some $k = 1, \ldots, n$, then follows (294) from (293).
- If $\varphi = \neg \psi$ and if the inductive hypothesis holds for $\psi$, then

$$\left( \psi[V \mapsto \Theta] \right)[v] = \left( \psi[V \mapsto \Theta] \right)[v] \overset{\text{ind}}{=} \psi[J] = \varphi[J].$$
If \( \varphi = \psi_1 \circ \psi_2, \circ \in \Sigma \) and if the inductive hypothesis holds for both \( \psi_1 \) and \( \psi_2 \), then

\[
\left( \psi[V \mapsto \Theta] \right)[v] = \left( \psi_1[V \mapsto \Theta] \right)[v] \circ \left( \psi_2[V \mapsto \Theta] \right)[v] \overset{\text{ind.}}{=} \psi_1[J] \circ \psi_2[J] = \varphi[J].
\]

In all cases, (294) holds.

**Proof of 840 (b).** The proof is identical to that of proposition 839 (a).

**Proof of 840 (c).** The proof is identical to that of proposition 839 (b) except for the special cases where renaming occurs, i.e. \( \varphi = \forall \xi \cdot \psi \) and \( \varphi = \exists \xi \cdot \psi \), where

- \( \xi \in \text{Free}(\mu) \).
- \( \eta \notin \text{Var}(\kappa) \cup \text{Var}(\mu) \cup \text{Var}(\psi) \).
- The inductive hypothesis holds for \( \psi \).

We will only show the case \( \varphi = \forall \xi \cdot \psi \) since the existential case is handled similarly. Since \( \xi \in \text{Free}(\mu) \), we have

\[
\varphi[\kappa \mapsto \mu] = \forall \eta \cdot \left( \psi[\xi \mapsto \eta][\kappa \mapsto \mu] \right),
\]

which does not allow us to use the inductive hypothesis directly.

We proceed to prove the statement by nested induction on the number of quantifiers. We have already shown the case of 0 quantifiers. Suppose that the statement holds for all formulas with strictly less than \( n \) quantifiers and suppose that \( \varphi \) has exactly \( n \) quantifiers.

Furthermore, for formulas with \( n \) quantifiers with \( \forall \) as the outermost one, the non-renaming cases (286) and (287) hold. Therefore, since \( \eta \notin \text{Free}(\mu) \),

\[
\varphi[\kappa \mapsto \mu][v] = \forall \eta \cdot \left( \psi[\xi \mapsto \eta][\kappa \mapsto \mu] \right)[v] \overset{(287)}{=} \left( \forall \eta \cdot \left( \psi[\xi \mapsto \eta] \right)[\kappa \mapsto \mu] \right)[v] \overset{\text{ind}}{=} \left( \forall \eta \cdot \left( \psi[\xi \mapsto \eta] \right) \right)[v] = \varphi[J],
\]

where we have implicitly used that \( \psi \) has \( n - 1 \) quantifiers.

On the other hand, due to proposition 839 (b),

\[
\left( \psi[\xi \mapsto \eta] \right)[v] = \psi[v_{\xi \mapsto \eta}]
\]

and, in particular, for any \( x \in X \),

\[
\left( \psi[\xi \mapsto \eta] \right)[v_{\eta \mapsto x}] = \psi[v_{\xi \mapsto \eta, \eta \mapsto x}] = \psi[v_{\xi \mapsto x}],
\]

422
where the last equality holds because $\eta \not\in \text{Var}(\psi)$.

Hence,
\[
\left( \forall \eta . (\psi[\xi \mapsto \eta]) \right)[v] = \frac{\left( \forall \xi . \psi \right)[v_{\xi \mapsto \eta}]}{\varphi[v_{\xi \mapsto \eta}]}.
\]

This proves (298).

\[\blacksquare\]

**Remark 841.** As explained in remark 825 (e), we avoid adding to a language more predicates than necessary. For this reason, we sometimes use predicate formulas. For example, if $\leq$ is a partial order symbol and we want to have a predicate for whether $\xi$ is the bottom element, we can define the formula

\[
\text{IsBottom}[\xi] := \forall \eta . \left( \xi \leq \eta \right).
\]

Note that $[\xi]$ is only a notational convenience for highlighting which variables are free, the actual formula is named \text{IsBottom}. This is consistent with remark 829 which allows us to write $\text{IsBottom}[\eta]$ rather than $\text{IsBottom}[\xi \mapsto \eta]$ to verify if $\eta$ is a bottom element.

**Proposition 842.** For any formula $\varphi$ and any variable $\xi$ over $\mathcal{L}$, we have the following equivalences:

\[
\neg \forall \xi . \varphi \Leftrightarrow \exists \xi . \neg \varphi \quad (301)
\]

\[
\neg \exists \xi . \varphi \Leftrightarrow \forall \xi . \neg \varphi \quad (302)
\]

**Proof.** The two equivalences are connected using double negation. We will only prove (301). Let $\mathcal{X} = (X, I)$ be a structure over $\mathcal{L}$ and let $v$ be a variable assignment. Then

\[
\left( \neg \forall \xi . \varphi \right)[v] = \left( \forall \xi . \varphi \right)[v] = \left( \bigwedge \{ \varphi[x \mapsto x] \mid x \in X \} \right)
\]

\[
= \bigvee \{ \varphi[x \mapsto x] \mid x \in X \} = \bigvee \{ (\neg \varphi)[x \mapsto x] \mid x \in X \} = (\exists \xi . \neg \varphi)[v].
\]

\[\blacksquare\]

**Proposition 843.** For any formula $\varphi$ and any variable $\xi$ over $\mathcal{L}$, the formulas $\varphi$ and $\forall \xi . \varphi$ are semantically equivalent.

This allows us to skip quantifiers when writing formulas without changing their validity. Given a formula $\varphi$ with free variables $\xi_1, \ldots, \xi_n$, we call

\[
\forall \xi_1 \cdot \ldots \cdot \forall \xi_n . \varphi
\]

its universal closure and say that $\varphi$ itself is implicitly universally quantified. Universal closures of quantifierless formulas are called universal formulas.

See example 902 (a) for how this fails for derivability.
Proof. Let $\mathcal{X} = (X, I)$ be a structure that satisfies $\varphi$. Let $v$ be a variable assignment in $\mathcal{X}$. Then for any $x \in X$, the modified variable assignment $u_{\xi \to x}$ also satisfies $\varphi$, i.e.

$$\varphi[v] = \varphi[u_{\xi \to x}] = T.$$ 

Thus, $\mathcal{X}$ is also a model for $\forall \xi . \varphi$.

Conversely, suppose that $\mathcal{X}$ satisfies $\forall \xi . \varphi$ and $v$ is any variable assignment. Then

$$\varphi[u_{\xi \to x}] = T$$

for any $x$, including $x := v(\xi)$. Thus,

$$\varphi[u_{\xi \to v(\xi)}] = \varphi[v] = T.$$ 

Therefore, $\mathcal{X}$ is also a model for $\varphi$. 

Proposition 844. Let $\mathcal{L}$ be a first-order language, $\varphi$ be a formula, $\xi$ be any variable and $\tau$ be a ground term in $\mathcal{L}$. The following hold:

(a) $\forall \xi . \varphi \models [\xi \mapsto \tau]$, 

(b) $\varphi[\xi \mapsto \tau] \models \exists \xi . \varphi$.

See also definition 901 (b) for inference rules corresponding to this proposition.

Proof. The proof is very straightforward, but the technical details make it look a bit more complicated.

First note that if the formulas on the left are unsatisfiable, the proof is trivial. Hence, we will assume that they are satisfiable.

Proof of 844 (a). From proposition 843 it follows that $\forall \xi . \varphi \models \varphi$. If $\xi$ is not free in $\varphi$, then $\varphi[\xi \mapsto \tau] = \varphi$ and the proof is finished. Suppose that $\xi$ is free in $\varphi$.

Let $\mathcal{X} = (X, I)$ be a model of $\forall \xi . \varphi$. Let $v$ be a variable assignment in $\mathcal{X}$ and let $t := \tau[v]$. To avoid the case where $\xi \in \text{Var}(\tau)$, we replace it with $\eta$ that is not a variable in neither $\tau$ nor $\varphi$. Then

$$\varphi[\xi \mapsto \tau][v] = \varphi[\xi \mapsto \eta][\eta \mapsto \tau][v] = \varphi[\xi \mapsto \eta][\eta \mapsto \tau][u_{\eta \mapsto \tau}] \overset{840 (c)}{=} \varphi[\xi \mapsto \eta][u_{\eta \mapsto \tau}] \overset{839 (b)}{=} \varphi[u_{\xi \to \eta \mapsto \tau}] = \varphi[u_{\xi \to \tau}].$$

Since $(\forall \xi . \varphi)[v] = T$, by definition of valuation of $\forall$ we have

$$\varphi[u_{\xi \to \tau}] = T.$$ 

Therefore, $\varphi[\xi \mapsto \tau][v] = T$ and, since $v$ was chosen arbitrarily, we have $\mathcal{X} \models \varphi[\xi \mapsto \tau]$. 

424
Proof of 844 (b). For any model \( \mathcal{X} = (X, I) \) of \( \varphi \), any assignment \( v \) satisfies \( \varphi[\xi \mapsto \tau] \). Thus, \( \varphi[v_{\xi \mapsto \tau} \mid t := \tau[v]] = T \) for \( t := \tau[v] \) and hence \( v \) satisfies \( \exists \xi . \varphi \). Therefore, \( \mathcal{X} \) is also a model of \( \exists \xi . \varphi \).

Proposition 845. Let \( \varphi \) be a formula in the language \( \mathcal{L} \). If \( \tau = c \) for some constant \( c \) that does not occur in \( \varphi \), the formulas \( \varphi[\xi \mapsto c] \) and \( \exists \xi . \varphi \) are equisatisfiable.

If there is no such constant in \( \mathcal{L} \), we can instead define the extension \( \mathcal{L}' \) of \( \mathcal{L} \) by adjoining a new constant \( c \).

This general procedure of creating a new extension language in order to obtain an equisatisfiable formula without the outer quantifier is called existential quantifier elimination.

Proof. Proposition 844 (b) implies that \( \exists \xi . \varphi \models \varphi[\xi \mapsto c] \), which is even stronger than the statement that if \( \exists \xi . \varphi \) is satisfiable, so it \( \varphi[\xi \mapsto c] \).

For the other direction, suppose that \( \mathcal{X} = (X, I) \) satisfies \( \exists \xi . \varphi \). Fix a variable assignment \( v \). Then there exists a value \( x \) such that

\[
\varphi[v_{\xi \mapsto x} \mid t := \tau] = T. \tag{303}
\]

Define the interpretation

\[
\bar{I}(a) := \begin{cases} 
  x, & a = c, \\
  a, & a \in \text{Func}, \\
  a, & a \in \text{Pred}
\end{cases}
\]

as \( I \) modified at \( c \), so that \( c[v] = x \). It remains to show that the structure \( (X, \bar{I}) \) is a model of \( \varphi[\xi \mapsto \tau] \). Since the new structure has the same domain, \( v \) is a variable assignment in this new structure. Nevertheless, we will denote it by \( \bar{v} \) in order to distinguish between valuations in the two structures.

Since \( \xi[v] = c[v] = x \), from proposition 840 (c) it follows that

\[
\varphi[\xi \mapsto c] \mid v] = \varphi[v_{\xi \mapsto x}] \models T. \tag{303}
\]

Since the above holds for any assignment \( v \), the structure \( (X, \bar{I}) \) is a model of \( \varphi[\xi \mapsto c] \).

[Aut20] thm. 16.29

Theorem 846 (First-order semantic deduction theorem). Let \( \Gamma \) be a set of formulas over some first-order language, let \( \psi \) be an arbitrary formula and let \( \varphi \) be a closed formula.

Then the entailment \( \Gamma, \varphi \models \psi \) holds if and only if \( \Gamma \models \varphi \rightarrow \psi \) holds.

See remark 847 for the importance of the condition that \( \varphi \) is a closed formula.

Due to remark 833, this theorem also holds for propositional formulas.

Compare this result with theorem 882 (Syntactic deduction theorem).

Proof.

Proof of sufficiency. Let \( \Gamma, \varphi \models \psi \) and let \( \mathcal{X} = (X, I) \) be a model for \( \Gamma \).

- If \( \mathcal{X} \models \varphi \), then \( \mathcal{X} \models \Gamma \cup \{ \varphi \} \) and from our assumption we conclude \( \mathcal{X} \models \psi \). Hence, for any variable assignment \( v \) we have

\[
(\varphi \rightarrow \psi)[v] = (\varphi[v] \rightarrow \psi[v]) = (T \rightarrow T) = T.
\]

425
If $\mathcal{X} \not\models \varphi$, then, since $\varphi$ is closed and the valuation $\varphi[v]$ does not depend on the variable assignment $v$, for any $v$ we have

$$(\varphi \to \psi)[v] = (\varphi[v] \to \psi[v]) = (F \to \psi[v]) = T. \quad (304)$$

In both cases we conclude that $\mathcal{X} \models \varphi \to \psi$. Therefore, $\Gamma \models \varphi \to \psi$.

**Proof of necessity.** Let $\Gamma \models \varphi \to \psi$ and let $\mathcal{X} = (X, I)$ be a model for $\Gamma \cup \{\varphi\}$. Let $v$ be an arbitrary assignment in $\mathcal{X}$.

Obviously $\mathcal{X} \models \Gamma$ and $\mathcal{X} \models \varphi$. We thus have $(\varphi \to \psi)[v] = T$ and $\varphi[v] = T$, which only leaves $\psi[v] = T$ as a possible option.

Hence, $\mathcal{X} \models \psi$ and, consequently, $\Gamma \cup \{\varphi\} \models \psi$. \qed

**Remark 847.** In order to highlight the importance of closed formulas in certain theorems, we will take a close look at the proof of theorem 846 (First-order semantic deduction theorem).

Note that (304) only holds because the formula $\varphi$ is closed. If it were not closed, we could only conclude that there exists a variable assignment $v_0$ such that $\varphi[v_0] = F$. Clearly then $(\varphi \to \psi)[v_0] = T$. But there may exist another assignment $v$ such that $\varphi[v] = T$ and $\psi[v] = F$, which would imply that $\mathcal{X} \not\models \varphi \to \psi$. 

426
12.6. First-order models

Much of section 12.4 (First-order logic) is dedicated to semantic equivalences between logical formulas, which are formulated and proved using structures. This section is dedicated to the study of structures themselves and relations between them. While model theory is a wide topic, for the purposes of this document we are only interested in the following questions:

- Which subsets of a structure form a substructure?
  This is answered by definition 848 and by definition 852. Vacuously, if the language contains no functional symbols, by proposition 851 (a) every subset of a structure is a substructure. Such is the case with sets themselves, with partially ordered sets or with metric and topological spaces.
  Proposition 855 shows that the set of all substructures of a structure is worth studying in itself.

- Given a model of some set \( \Gamma \) of formulas, which substructures and homomorphic images of the model are again models of \( \Gamma \)?
  This is answered by proposition 863, proposition 865 and proposition 866.

**Definition 848.** Let \( \mathcal{X} = (X, I) \) be a structure for the language \( \mathcal{L} \) and let \( Y \subseteq X \). We say that \( Y = (Y, I) \) is a substructure of \( \mathcal{X} \) if it satisfies any of the following equivalent conditions:

(a) If \( Y \) is closed under function application, that is, for any functional symbol \( f \) in \( \mathcal{L} \) with arity \( n \), we have \( I(f)(Y^n) \subseteq Y \).

(b) The universe \( Y \) is a fixed point of the operator

\[
T : \text{pow}(X) \to \text{pow}(X)
\]

\[
T(A) := A \cup \{ x \in X \mid \exists f \in \text{Fun} \cdot \exists x_1, \ldots, x_{\#f} \in A \cdot f[x_1, \ldots, x_{\#f}] = x \},
\]  

(305)

which enlarges \( A \) with the union of all image of \( A \) under functions of the language \( \mathcal{L} \). Note that the formula inside (305) is in the metalanguage despite using syntax similar to first-order logic formulas.

**Proof.** By definition of \( T \), \( Y \) if a fixed point if and only if

\[
\{ x \in X \mid \exists f \in \text{Fun} \cdot \exists x_1, \ldots, x_{\#f} \in A \cdot f[x_1, \ldots, x_{\#f}] = x \} \subseteq Y.
\]

This condition is clearly satisfied if \( B \) satisfies definition 848 (a).

If, instead \( Y \) is a fixed point of \( T \), for the \( n \)-ary functional symbol \( f \in \text{Fun} \) and for any tuple \( x_1, \ldots, x_n \), the value \( I(f)(x_1, \ldots, x_n) \) belongs to \( Y \). Therefore, definition 848 (a) is satisfied.

**Remark 849.** Let \( \mathcal{X} = (X, I) \) be a structure over some language \( \mathcal{L} \) without predicate symbols. If, for every functional symbol \( f \), the interpretation \( I(f) \) is a continuous function, we call \( \mathcal{X} \) a topological structure.
For every algebraic structure defined in section 9 (Group theory) and section 10 (Ring theory), there exists a topological equivalent. We discuss topological groups and topological vector spaces through the document, especially in section 4 (Functional analysis).

Naturally, every substructure of a topological structure is again a topological structure.

Example 850. The classic definition for a subset $U$ of a vector space $V$ being a vector subspace is that $U$ is closed under linear combinations. Linear combinations are simply finitesuperpositions ofaddition and scalar multiplication in $V$. So this condition ensures that $U$ is closed under application of the functional symbols corresponding to addition and scalar multiplication.

See proposition 554 for a further discussion.

Proposition 851. Fix a language $\mathcal{L}$. The first order substructures of a structure $\mathcal{X} = (X, I)$ have the following basic properties:

(a) If $\mathcal{L}$ has no functional symbols, then $(Y, I)$ is a substructure where $Y$ is any subset of $X$.

(b) Let $\{(Y_k, I)\}_{k \in \mathcal{K}}$ be a family of substructures of $\mathcal{X}$. Then their intersection structure $(\bigcap_{k \in \mathcal{K}} Y_k, I)$ is again a substructure of $\mathcal{X}$.

Proof.
Proof of 851 (a). Both conditions in definition 848 are vacuously satisfied if there are no functional symbols in $\mathcal{L}$.

Proof of 851 (b). For any functional symbol $f$ in $\mathcal{L}$ with arity $n$, we have

$$I(f)\left(\bigcap_{k \in \mathcal{X}} Y_k\right)^n \subseteq \bigcap_{k \in \mathcal{X}} I(f)(Y_k^n) \subseteq \bigcap_{k \in \mathcal{X}} Y_k.$$

Therefore, $(\bigcap_{k \in \mathcal{X}} Y_k, I)$ is indeed a substructure of $\mathcal{X}$.

Definition 852. Let $\mathcal{X} = (X, I)$ be a structure over $\mathcal{L}$ and let $A \subseteq X$ be any set. The set $A$ is said to generate\(^{13}\) the substructure $\mathcal{Y} = (Y, I)$ if it satisfies any of the equivalent statements:

(a) Out of all substructures of $\mathcal{X}$ whose domain contains $A$, the domain $\mathcal{Y}$ is the smallest with respect to set inclusion.

(b) $\mathcal{Y}$ is the intersection structure of all substructures of $\mathcal{X}$ that contain $A$.

Proof. Let $\{(Y_k, I)\}_{k \in \mathcal{K}}$ be the family of all substructures of $\mathcal{X}$ whose domains contain $A$. Fix one of these substructures, say $(Y_{k_0}, I)$.

We have the obvious inclusion

$$\bigcap_{k \in \mathcal{K}} Y_k \subseteq Y_{k_0}.$$

The reverse inclusion holds if and only if $Y_{k_0}$ is contained in each one of domains $Y_k$ for $k \in \mathcal{K}$. In other words, $Y_{k_0}$ is the smallest of the domains $\{Y_k\}_{k \in \mathcal{K}}$ with respect to set inclusion if and only if $Y_{k_0}$ equals their intersection.\(^{13}\)

\(^{13}\)bg: поражда, ru: порождает
**Example 853.** Common examples of generated substructures are the linear span discussed in proposition 554 and the generated ring ideals.

**Proposition 854.** Let \( \mathcal{X} = (X, I) \) be a structure for the language \( \mathcal{L} \). Every subset of \( X \) has a generated substructure.

*Proof.* Given a set \( Y \subseteq X \), we apply theorem 1259 (Knaster-Tarski theorem) to the Boolean algebra of all subsets \( \text{pow}(X) \) with the operator

\[
R : \text{pow}(X) \to \text{pow}(X)
\]

\[
R(A) := Y \cup T(A),
\]

where \( T \) is defined in (305).

We thus obtain the smallest fixed point \( Z \) of \( T \), which contains \( Y \) and satisfies definition 848 (b). The structure \( \mathcal{Z} = (Z, I) \) is thus a substructure of \( \mathcal{X} \).

Furthermore, \( \mathcal{Z} \) is unique because its domain is the smallest set invariant under \( R \). \( \square \)

**Proposition 855.** Fix a structure \( \mathcal{X} = (X, I) \) for the language \( \mathcal{L} \).

The set of all substructures of a \( \mathcal{X} \) forms a complete lattice with respect to set inclusion of domains. It is isomorphic to a complete sublattice of the Boolean algebra \( \text{pow}(X) \) described in proposition 938.

Explicitly:

(a) The join of the family of substructures with domains \( \{ (Y_k, I) \}_{k \in \mathcal{X}} \) is the generated substructure of the set \( \bigcup_{k \in \mathcal{X}} Y_k \).

(b) The top element is the substructure \( \mathcal{X} \) itself. Any substructures that are different from \( \mathcal{X} \) are called proper.

(c) The meet of the family of substructures \( \{ (Y_k, I) \}_{k \in \mathcal{X}} \) is simply the intersection structure of the family.

(d) The bottom element of this lattice is the intersection of all substructures. This is called the trivial substructure. As a matter of fact, because the trivial substructures for any two structures are isomorphic, we refer to them collectively as the trivial structure because trivial substructures are isomorphic.

As discussed in remark 827, the empty set is not allowed to be the domain of a structure by definition. Nevertheless, for the sake of having a bottom element we allow structures with empty domains in this lattice.

By proposition 1187 (b), the trivial structures are the initial objects of all categories of models.

The trivial substructure usually consists only of the constants of \( \mathcal{L} \) — for example, the trivial group \( \{ e \} \) or the trivial bounded lattice \( \{ T, \bot \} \).

*Proof.*
Proof of 855 (a). Let \((Y, I)\) be the generated substructure of the set \(A := \bigcup_{k \in \mathcal{K}} Y_k\). From definition 852 (a) it follows that out of the domains of all substructures of \(\mathcal{X}\), \(Y\) is the smallest that contains \(A\) and hence the smallest that contains \(Y_k\) for all \(k \in \mathcal{K}\). Therefore, it is indeed the supremum of the family \(\{(Y_k, I)\}_{k \in \mathcal{K}}\) with respect to set inclusion of domains.

Proof of 855 (b). Since \(\mathcal{X}\) is a substructure of itself, it is not only the supremum of the entire lattice, but actually the maximum.

Proof of 855 (c). The domain of the intersection structure of the family \(\{(Y_k, I)\}_{k \in \mathcal{K}}\) of substructures of \(\mathcal{X}\) is obviously the infimum of the family since its domain are the infimum in the Boolean algebra of subsets of \(X\).

Proof of 855 (d). It trivially follows from proposition 855 (c) that the bottom element is the intersection of all substructures of \(\mathcal{X}\).

Definition 856. Let \(\mathcal{X} = (X, I)\) and \(\mathcal{Y} = (Y, I)\) be structures over a common language. We say that the function \(h : X \to Y\) is a homomorphism between \(\mathcal{X}\) and \(\mathcal{Y}\) if it preserves all functions and relations. Explicitly:

(a) For any functional symbol \(f \in \text{Fun}\) of arity \(n\) and any tuple \(x_1, \ldots, x_n \in X\) we have
\[
h(I_X(f)(x_1, \ldots, x_n)) = I_Y(f)(h(x_1), \ldots, h(x_n))
\]

(b) For any predicate symbol \(p \in \text{Pred}\) of arity \(n\) and any \(x_1, \ldots, x_n \in X\),
\[
I_X(p)(x_1, \ldots, x_n) = T \implies I_Y(p)(h(x_1), \ldots, h(x_n)) = T.
\]

Remark 857. Homomorphisms as they are defined in definition 856 are sometimes called weak homomorphisms. Under weak homomorphisms, it is possible that \(I_X(p)(x_1, \ldots, x_n) = F\) and yet \(I_Y(p)(h(x_1), \ldots, h(x_n)) = T\).

Logicians sometimes define strong homomorphisms where they replace definition 856 (b) with the stronger condition
\[
I_X(p)(x_1, \ldots, x_n) = I_Y(p)(h(x_1), \ldots, h(x_n)).
\]

This condition seems much more natural at first, but it is less useful in practice. For example, monotone maps and graph homomorphisms are both weak homomorphisms and these are the most used definitions of homomorphisms in languages with predicate symbols. For this reason, we mostly avoid studying strong homomorphisms.

They are useful for certain propositions like proposition 1243, however.

Proposition 858. First-order structure homomorphisms have the following basic properties:

(a) If \(\mathcal{X} = (X, I)\) is a structure and \(\mathcal{Y} = (Y, I)\) is a substructure of \(\mathcal{X}\), then the canonical embedding function
\[
t : Y \to X \quad \quad t(y) := y
\]
is indeed a homomorphism (and thus an embedding in the sense of definition 859).
(b) If $\mathcal{X} = (X, I_X)$ and $\mathcal{Y} = (Y, I_Y)$ are structures and $h : X \to Y$ is a (weak) homomorphism, then the image $h(\mathcal{X}) := (h(X), I_Y)$ is a substructure of $\mathcal{Y}$.

(c) The composition of two homomorphisms is again a homomorphism.

(d) Fix a homomorphism $h : X \to Y$ and a term $\tau$. For any variable assignments $v_X$ and $v_Y$ such that $v_Y(\xi) = h(v_X(\xi))$ for all $\xi \in \text{Var}(\tau)$, we have

$$h(\tau \downarrow v_X) = \tau \downarrow v_Y.$$  

Proof.

Proof of 858 (a). The interpretation in the substructure $\mathcal{Y}$ is the restriction $I|_Y$ of $I$ to $\mathcal{Y}$ which simply restricts the domain of any predicate and function is indeed an interpretation in $\mathcal{Y}$. Thus, $(\mathcal{Y}, I|_Y)$ is a structure.

Conditions definition 856 (a) and definition 856 (b) are both satisfied since the interpretation of any function and predicate is restricted to $\mathcal{Y}$. Thus, $\iota$ is a homomorphism.

Proof of 858 (b). To prove that $(f(X), I_Y)$ is a substructure of $\mathcal{Y}$, we will show that definition 848 (a) holds.

Indeed, due to definition 856 (a), for any $n$-ary functional symbol and any tuple $x_1, \ldots, x_n \in X$, we have that

$$I_Y(f)(h(x_1), \ldots, f(x_n)) \overset{856 (a)}{=} h(I_X(f)(x_1, \ldots, x_n)) \overset{848 (a)}{\in} h(X).$$

Proof of 858 (c). Let $h : \mathcal{X} \to \mathcal{Y}$ and $l : \mathcal{Y} \to \mathcal{Z}$ both be homomorphisms.

- Definition 856 (a) is satisfied because for any $n$-ary functional symbol $f$ and any tuple $x_1, \ldots, x_n \in X$,

$$ (l \circ h)(I_X(f)(x_1, \ldots, x_n)) \overset{856 (a)}{=} 
= l(I_Y(f)(h(x_1), \ldots, h(x_n))) \overset{856 (a)}{=} 
= I_Z(f)((l \circ h)(x_1), \ldots, (l \circ h)(x_n)).$$

- Definition 856 (b) is satisfied because for any $n$-ary predicate symbol $p$ and any tuple $x_1, \ldots, x_n \in X$,

$$ I_X(p)(x_1, \ldots, x_n) \overset{856 (b)}{=} 
= I_Y(p)(h(x_1), \ldots, h(x_n)) \overset{856 (b)}{=} 
= I_Z(p)((l \circ h)(x_1), \ldots, (l \circ h)(x_n)).$$
Proof of 858 (d). We use induction on the structure of \( \tau \). If \( \tau \) is a variable, the statement is obvious from the compatibility condition for \( v_X \) and \( v_Y \). If \( \tau = f(\kappa_1, \ldots, \kappa_m) \), then
\[
\tau [v_Y] = I(f)(\kappa_1 [v_Y], \ldots, \kappa_m [v_Y]) = \\
= I(f)(h(\kappa_1 [v_X], \ldots, h(\kappa_m [v_X])) \quad \text{by (a)} \\
= h(I(f)(\kappa_1 [v_X], \ldots, \kappa_m [v_X])) = \\
= h(\tau [v_X]).
\]

\( \square \)

Definition 859. In connection with definition 971 and definition 1125, and more importantly proposition 917, we introduce the following terminology for homomorphisms:

(a) An embedding is an injective homomorphism.

(b) The dual surjective homomorphisms do not have an established name. The term projection is often used, but projections are conventionally idempotent.

(c) An isomorphism is a bijective homomorphism whose inverse is also a homomorphism. Equivalently, it is a bijective strong homomorphism.

The peculiarity here is that the inverse in the sense of definition 1125 (e) may not be a homomorphism in general. See example 860. For examples where this condition can be relaxed, see proposition 861 and corollary 1244.

(d) An endomorphism is a homomorphism that is also an endofunction.

(e) A homomorphism that is both an endomorphism and an isomorphism is called an automorphism.

Example 860. Consider the set of integers \( \mathbb{Z} \) endowed with two different partial orders:

- The standard total order \( \leq \) where \( n \leq m \) if there exists a nonnegative integer \( k \) such that \( n + k = m \).
- The equality \( = \) relation, which is simply the diagonal relation \( \Delta \).

The identity \( \text{id}(x) = x \) is an order homomorphisms from \((\mathbb{Z}, =)\) to \((\mathbb{Z}, \leq)\). Indeed, for any integers \( n \) and \( m \), \( n = m \) implies \( n \leq m \).

Furthermore, the identity function is bijective. The inverse of \( \text{id} \), which is again \( \text{id} \), is not however a homomorphism from \((\mathbb{Z}, \leq)\) to \((\mathbb{Z}, =)\) since, for example, \( 1 \leq 2 \), but \( 1 \neq 2 \).

Hence, \( \text{id} : (\mathbb{Z}, =) \to (\mathbb{Z}, \leq) \) is a bijective homomorphism, but not an isomorphism.

Proposition 861. Let \( \mathcal{L} \) be a language without predicate symbols. Then any bijective homomorphism between structures for \( \mathcal{L} \) is an isomorphism.

This applies to arbitrary languages if we restrict ourselves to strong homomorphisms.
Proof. If $\mathcal{L}$ has no predicate symbols, then definition 856 (b) is vacuously satisfied for the inverse of any bijective homomorphism. It is just as well satisfied if we restrict ourselves to strong homomorphisms. \hfill $\square$

**Definition 862.** We say that a propositional formula $\varphi$ is **positive** if it contains only positive literals connected using conjunction $\land$ and disjunction $\lor$. We can also add propositional constants, however that would be redundant due to proposition 1255 (b).

The point of positive formulas is to avoid negation $\neg$. This definition is not equivalent to positive implicational formulas where $\rightarrow$ is the only connective. We avoid adding $\rightarrow$ because that would allow us, assuming classical logic, to derive negation using proposition 815 (a).

Positive formulas are used in proposition 863, which fails to hold for some non-positive formulas — see example 449.

When dealing with first-order logic, we simply use substitution to replace propositional variables with atomic formulas. This way we obtain positive first-order formulas with implicit universal quantification. Of course, we can always add explicit universal quantifiers, but we avoid existential quantifiers because of proposition 842.

**Proposition 863.** Let $\mathcal{X} = (X, I)$ and $\mathcal{Y} = (Y, I)$ be structures over a common language $\mathcal{L}$ and let $h : X \rightarrow Y$ be a homomorphism between them. Take $\Gamma$ to be a nonempty set of positive formulas.

Then $h$ preserves models of $\Gamma$. That is, if $\mathcal{X} \models \Gamma$ then $(h(X), I_Y) \models \Gamma$.

**Proof.** Let $v$ a variable assignment in $\mathcal{Y}$. Let $v_X : \text{Var} \rightarrow X$ be an assignment such that for any variable $\xi$ we have

$$v_X(\xi) \in h^{-1}(v_Y(\xi)).$$

At least one such assignment exists due to definition 984 (g). If $h$ is injective, this assignment is unique.

We will show that

$$\mathcal{X} \models_{v_X} \varphi$$

implies

$$(h(X), I_Y) \models_{v_Y} \varphi.$$  \hfill (307)

We assume (307) for $\varphi$ and we use induction on the structure of $\varphi$ to prove (308), starting with different atomic formulas:

- The constant $\top$ is vacuously preserved by homomorphisms because it does not depend on the interpretation or variable assignment.

- Suppose that $\varphi = (\tau_1 \equiv \tau_2)$. We have $\tau_1 \|$ and $\tau_2 \|$ and hence $h(\tau_1 \| ) = h(\tau_2 \| )$ and

$$\tau_1 \| = h(\tau_1 \| ) = h(\tau_2 \| ) = \tau_2 \|.$$

- Suppose that $\varphi$ is the predicate formula $p(\tau_1, \ldots, \tau_n)$. By assumption for every variable assignment in $\mathcal{X}$ and, in particular, for any $v_X$,

$$\mathcal{X} \models_{v_X} p(\tau_1, \ldots, \tau_n).$$

433
then
\[ I_X(p)(\tau_1[v_X], ..., \tau_n[v_X]) = T. \] (309)

By definition of homomorphism, we have
\[ I_Y(p)(h(\tau_1[v_X]), ..., h(\tau_n[v_X])) = T. \] (310)

Now
\[ (h(X), I_Y) \models_{v_Y} p(\tau_1, ..., \tau_n), \]
follows from proposition 858 (d).

If \( h \) is a strong homomorphism, then the converse also holds, i.e. (309) follows from (310). See proposition 865 for an application of this converse.

- Suppose that \( \varphi = \psi_1 \land \psi_2 \) and that the inductive hypothesis holds for \( \psi_1 \) and \( \psi_2 \).
  
  Since \( \varphi[v_X] = T \) by assumption, by definition of valuation of conjunction we have
  \[ \psi_1[v_X] = \psi_2[v_X] = T. \]
  
  This allows us to apply the inductive hypothesis to obtain
  \[ \psi_1[v_Y] = \psi_2[v_Y] = T. \]
  
  and conclude that
  \[ \varphi[v_Y] = \psi_1[v_Y] \land \psi_2[v_Y] = T \land T = T. \]

- Suppose that \( \varphi = \psi_1 \lor \psi_2 \) and that the inductive hypothesis holds for \( \psi_1 \) and \( \psi_2 \).
  
  Since the formula \( \varphi \) is valid in \( \mathcal{X} \), at least one of \( \psi_1 \) or \( \psi_2 \) is valid under \( v_X \). For different \( v_X \) the valuation pair \( (\psi_1[v_X], \psi_2[v_X]) \) may be different, but will always have at least one \( T \) value.
  
  The inductive hypothesis holds for both \( \psi_1 \) and \( \psi_2 \) and therefore \( (\psi_1[v_Y], \psi_2[v_Y]) \) also contains at least one \( T \) value.
  
  This allows us to conclude that
  \[ \varphi[v_Y] = \psi_1[v_Y] \lor \psi_2[v_Y] = T. \]

- To see how this proof fails for conditionals, consider \( \varphi = (\psi_1 \rightarrow \psi_2) \). Then \( \varphi[v_X] = T \) implies either \( \psi_1[v_X] = F \) or \( \psi_1[v_X] = \psi_2[v_X] = T. \)
  
  If \( \psi_1[v_X] = \psi_2[v_X] = F, \) we have \( \varphi[v_X] = T, \) but we cannot conclude that \( \varphi[v_Y] = T \) because that would require the inverse of the inductive hypothesis to hold for \( \psi_1 \) and \( \psi_2 \).
  
  See example 449 for an example where a conditional is not preserved by a homomorphism.

434
Since \( v \) was chosen arbitrarily, we conclude that
\[
(h(X), I_Y) \models \varphi.
\]

\[ \square \]

**Corollary 864.** If \( \Gamma \) is a set of positive formulas, any substructure of a model of \( \Gamma \) is again a model of \( \Gamma \).

**Proof.** Follows from proposition 858 (a) and proposition 863.

**Proposition 865.** If \( \mathcal{X} = (X, I_X) \) is a model of \( \Gamma \) and if \( h : X \to Y \) is a embedding from \( \mathcal{X} \) to \( Y = (Y, I_Y) \), then \( (h(X), I_Y) \) is also a model of \( \Gamma \).

We say that embeddings preserve arbitrary formulas.

If \( (h(X), I_Y) \) is a model of \( \Gamma \), then \( \mathcal{X} \) is also a model if \( h \) is a strong homomorphism.

We say that strong embeddings reflect arbitrary formulas.

**Proof.** The proof simply extends the induction in the proof of proposition 865 to
\[
\varphi[v_Y] = \varphi[v_Y]
\]
which allows us to use the usual induction on the negation and all connectives and quantifiers.

The result regarding strong homomorphisms is shown in the note about (309) following from (310) under strong homomorphisms.

**Proposition 866.** Let \( \Gamma \) to be a nonempty set of positive formulas. Let \( \mathcal{X} \) be a model of \( \Gamma \) and let \( A \) be any nonempty set. Consider the set \( Y := \text{Fun}(A, \mathcal{X}) \) of all set functions from \( A \) to \( X \).

Define the function \( \iota : X \to Y \) by sending each \( x \in X \) to the corresponding constant function in \( Y \).

Define the interpretation \( I_Y \) as follows:

- For each \( n \)-ary functional symbol \( f \) in \( \mathcal{L} \), define the interpretation of the functions \( k_1, \ldots, k_n \) componentwise as
  \[
  I_Y(f) : Y^n \to Y
  \]
  \[
  I_Y(f)(k_1, \ldots, k_n) := \left( s \mapsto I(f)(k_1(s), \ldots, k_n(s)) \right).
  \]

- For each \( n \)-ary predicate symbol \( p \) in \( \mathcal{L} \), define \( I_Y(p) \subseteq Y^n \) via
  \[
  I_Y(p) : Y^n \to \{T, F\}
  \]
  \[
  I_Y(p)(k_1, \ldots, k_n) := \bigwedge \{ I(p)(k_1(s), \ldots, k_n(s)) \mid s \in S \}.
  \]

That is, \( I_Y(p)(k_1, \ldots, k_n) = T \) if and only if \( I(p)(k_1(s), \ldots, k_n(s)) = T \) simultaneously for all \( s \in S \).

Then the structure \( Y = (Y, I_Y) \) is also a model of \( \Gamma \) and \( \iota : \mathcal{X} \to Y \) is an embedding.
Proof. It is obvious that \( Y \) is a structure and that \( \iota \) is an embedding. We will prove using induction on the structure of a formula \( \varphi \) that \( X \models \varphi \) implies \( Y \models \varphi \).

Let \( \nu_Y \) be a variable assignment in \( Y \). Suppose that \( X \models \varphi \). We use induction on the structure of \( \varphi \) to show that \( \varphi[\nu_Y] \) is valid.

- If \( \varphi \) is a propositional constant, its value does not depend on \( \nu_Y \) and thus \( \varphi[\nu_Y] = \varphi[\nu_X] \).

- If \( \varphi = (\tau_1 \equiv \tau_2) \), then \( \tau_1[\nu_X] = \tau_2[\nu_X] \) for all assignments \( \nu_X \) in \( X \), hence for any \( s \in S \) we have \( (\tau_1[\nu_Y])(s) = (\tau_2[\nu_Y])(s) \) since both sides of the equality here are elements of \( X \).

- Analogously, if \( \varphi = p(\tau_1, \ldots, \tau_n) \), then
  \[
  I_Y(p)(k_1, \ldots, k_n) = \bigwedge \left\{ I(p)(k_1(s), \ldots, k_n(s)) \mid s \in S \right\} = \bigwedge \{ T \mid s \in S \} = T.
  \]

- Analogous to the proof of proposition 863, conjunction and disjunction formulas that are valid in \( X \) are valid in \( Y \) while conditional formulas may fail to be valid. See example 867 for examples where this proposition fails.

Example 867. While the statement of proposition 866 may be a little cryptic, a few examples show that it is actually obvious.

- Boolean functions have their values in the Boolean algebra \( \{ T, F \} \). Let \( S \) be the set of all tuples of values in \( \{ T, F \}^n \) for arbitrary \( n \). That is,
  \[
  S := \bigcup_{n \geq 1} \{ T, F \}^n.
  \]
  Then from proposition 866 it follows that the set \( B = \text{fun}(S, \{ T, F \}) \) of all Boolean functions of arbitrary arities is again a Boolean algebra. See theorem 814 (b) for further discussion.

- If \( R \) is a ring and \( A \) is any set, then \( \text{fun}(A, R) \) is again a ring with componentwise operations — see proposition 568.
  This is useful in functional analysis where we study real-valued and complex-valued functions over arbitrary sets.

- If \( \mathbb{K} \) is a field, then in general \( \text{fun}(A, \mathbb{K}) \) is not a field. The simplest example are the real-valued real functions — \( \sin(x) \) has no multiplicative inverse since \( 1/\sin(x) \) is not defined for \( x = 2k\pi, k = 1, 2, \ldots \). We can form a field of fractions, but in general fields of fractions over function rings no longer correspond to functions — they are purely algebraic constructions, just like formal power series.

436
This happens because the definition of a field and, more generally has a non-positive axiom — it requires every nonzero element to have a multiplicative inverse, which can be described formally as

\[
\forall \xi \cdot ((\xi \neq 0) \lor \exists \eta \cdot (\xi \cdot \eta = 1)).
\]

As discussed in definition 862, a formula with an existential quantifier may fail to be positive.

**Definition 868.** Fix a first-order language \( \mathcal{L} \) and a structure \( \mathcal{X} = (X, I) \) for \( \mathcal{L} \). To every formula formula \( \varphi[\xi_1, \ldots, \xi_n] \) there corresponds a set \( A \subseteq X^n \) such that

\[(x_1, \ldots, x_n) \in A \text{ if and only if } \varphi[x_1, \ldots, x_n] = T.\]

We say that \( \varphi \) defines \( A \). An arbitrary set \( A \subseteq X^n \) is definable if there exists a formula \( \varphi \) that defines \( A \).

See theorem 1104 for how this concept deeply relates to set theory.

**Proposition 869.** Let \( \mathcal{X} = (X, I) \) be a structure over some language \( \mathcal{L} \). Then for every formula \( \varphi[\xi_1, \ldots, \xi_n] \) in \( \mathcal{L} \) and any automorphism \( h : X \to X \) we have

\[
\varphi[x_1, \ldots, x_n] = \varphi[h(x_1), \ldots, h(x_n)].
\]

**Proof.** Let \( x_1, \ldots, x_n \) be points in \( X \) such that

\[
\varphi[x_1, \ldots, x_n] = T.
\]

Since \( h \) is a homomorphism, it follows that

\[
\varphi[h(x_1), \ldots, h(x_n)] = \varphi[x_1, \ldots, x_n] = T.
\]

Conversely, suppose that \( x_1, \ldots, x_n \) are points in \( X \) such that

\[
\varphi[h(x_1), \ldots, h(x_n)] = T.
\]

Then since \( h^{-1} \) is a homomorphism, we have

\[
\varphi[h^{-1}(h(x_1)), \ldots, h^{-1}(h(x_n))] = \varphi[x_1, \ldots, x_n] = T.
\]

Therefore,

\[
\varphi[h(x_1), \ldots, h(x_n)] = T \text{ if and only if } \varphi[x_1, \ldots, x_n] = T,
\]

which is equivalent to the statement of the proposition.

**Corollary 870.** Let \( \mathcal{X} = (X, I) \) be a structure over some language \( \mathcal{L} \).

(a) If the set \( A \subseteq X \) is definable and if \( h : X \to X \) is an automorphism, then \( h(A) = A \).

(b) If for some automorphism \( h : X \to X \) we have \( h(A) \neq A \), then the set \( A \subseteq X \) is not definable.

**Proof.**

**Proof of 870 (a).** If \( A \) is definable via \( \varphi[\xi] \), then by proposition 869 for any automorphism \( h : X \to X \) we have \( \varphi[x] = \varphi[h(x)] \). Thus, the set \( h(A) \) is just as well-definable via \( \varphi \), which implies \( A = h(A) \).

**Proof of 870 (b).** This is the contrapositive of corollary 870 (a).
12.7. Deductive systems

Without a clear context, by “logical formula” we will mean a word over some alphabet. In practice, these will be either propositional formulas or first-order formulas.

It is challenging to formally define a deductive system in a way that reflects reality. We will do this via some auxiliary definitions that rely heavily on the interaction between the object logic and the metalogic.

**Definition 871.** A **judgment** is a logical formula of the metalogic, which is usually used for assertion. Using judgments allows us to quantify over metalogical properties. We will be interested in the following kinds of judgments:

(a) A **sequent** is a judgment with at least two free variables. We denote a sequent $\Gamma \vdash \Delta$ with free variables $\Gamma$ and $\Delta$ by

$$\Gamma \vdash \Delta.$$  

The intended interpretation for both $\Gamma$ and $\Delta$ is that of sets of formulas in the object logic, in which case a sequent expresses the metalogical statement that “the formulas in $\Gamma$ collectively entail via $\vdash$ the any formula in $\Delta$”.

We could have defined a sequent as a predicate, however we may not have appropriate predicate symbols on the metalanguage. This issue is discussed in remark 841.

(b) An **inference rule** is a judgment with at least one free variable. We denote an inference rule $R$ with free variables $\psi$ and $\varphi_1, \ldots, \varphi_n$ by

$$\frac{\varphi_1 \ \cdots \ \varphi_n}{\psi} \quad R$$

We allow the possibility that $n = 0$, in which case the rule becomes

$$\frac{}{\psi} \quad R$$

The intended interpretation for the listed variables is that of formulas in the object logic, in which case an inference rule expresses the metalogical statement that “the premises $\varphi_1, \ldots, \varphi_n$ collectively entail the consequence $\psi$, as justified by the rule $R$”. When building complicated proofs, however, the applicability of the rule may depend on some context, and for this reason $\varphi_1, \ldots, \varphi_n$ are often interpreted as proof trees rather than single formulas. See definition 901 (a) for such an example.

**Remark 872.** Inference rules are both special cases and generalizations of sequents, depending on how we view them. Using an interpretation where $\Gamma = \varphi_1, \ldots, \varphi_n$ and $\Delta = \psi$, inference rules simply allow an alternative syntax for sequents.

It is sometimes convenient, however, to interpret the formulas $\varphi_1, \ldots, \varphi_n$ as sequents in some auxiliary logic between the object and metalogic, in which case we are able to express more complicated inference rules of the sort

$$\varphi, \Gamma \vdash \Delta, \psi$$

$$\frac{\varphi, \Gamma \vdash \Delta, \psi}{\Gamma \vdash \Delta, \varphi \rightarrow \psi}$$

438
This can be very useful when inductively defining the metalogical relation $\vdash$.

**Definition 873.** A **proof tree** is a rooted tree of logical formulas. The validity of a proof tree is handled by deductive system and is irrelevant for this definition.

(a) A **subproof** is simply a subtree of a proof. If a subproof was obtained using an inference rule, the subproof labeled with the name of the rule by the deductive systems.

(b) The root of the tree is called the **conclusion** of the proof and the leaves are called **premises** or, in the context of natural deduction, **assumptions**. We sometimes add a non-premise label to a leaf that prevents it from being added to the list of premises.

(c) We will draw graphically proof trees with no edges and with the root at the bottom in a style inspired by inference rules. See example 879 for an example.

**Definition 874.** A **deductive system** for a set $\mathcal{F}$ of formulas in the object logic is a metalogical collection of inference rules, which are used to generate proof trees in the object logic.

We will not attempt to encode inference rules themselves into the object theory and instead regard a deductive system as a pair $(\mathcal{F}, \mathcal{P})$, where $\mathcal{F}$ is a set of logical formulas and $\mathcal{P}$ is a set of proofs over $\mathcal{F}$.

We define the set $\mathcal{P}$ of proofs via structural recursion.

(a) For every formula $\varphi$ in $\mathcal{F}$, the tree with root $\varphi$ and no children is a proof $\mathcal{P}$.

(b) Let $\varphi_1, \ldots, \varphi_n$ and $\psi$ be formulas in $\mathcal{F}$ and $\mathcal{P}_1, \ldots, \mathcal{P}_n$ be proofs of $\varphi_1, \ldots, \varphi_n$.

Suppose that the deductive system has an inference rule

\[
\frac{\Phi_1 \quad \ldots \quad \Phi_n}{\Psi} \quad R
\]

such that

\[
R[\Phi_1 \mapsto \mathcal{P}_1, \ldots, \Phi_n \mapsto \mathcal{P}_n, \Psi \mapsto \psi] = T
\]

in the metalogic of the metalogic.

Then the tree with root $\psi$, subtrees $\mathcal{P}_1, \ldots, \mathcal{P}_n$ of the root, and label $R$, is a proof in $\mathcal{P}$. In the case of rules with no premises like $(\top^+)$, we add a non-premise label to the proof in order to exclude $\psi$ from the list of premises of the proof.

We say that the resulting proof

\[
\frac{\mathcal{P}_1 \quad \ldots \quad \mathcal{P}_n}{\psi} \quad r
\]

is an **application** of the rule $R$. 
Definition 875. We are often interested not in the proofs of a deductive system, but in \textit{provability}, which we express via \textit{sequents}.

Fix a deductive system \((\mathcal{F}, \mathcal{P})\). If \(\mathcal{P}\) contains a proof of \(\varphi\), whose premises are all members of \(\Gamma\), the following sequent is valid:

\[ \Gamma \vdash \varphi. \]

We say that \(\varphi\) is a \textit{theorem} of \(\Gamma\). If \(\Gamma\) is empty, we say that \(\varphi\) is a \textit{logical theorem}.

Furthermore, \(\vdash\) is a reflexive and transitive relation, which makes \((\text{pow}(\mathcal{F}), \vdash)\) a \textit{preordered set}.

Definition 876. \textbf{Axiomatic deductive systems}, also called \textit{Hilbert-style systems}, are \textit{Author’s definition} deductive systems for propositional formulas consist of the single \textit{inference rule} \textit{modus ponens}\textsuperscript{14}:

\[
\begin{array}{c}
\varphi \\
\varphi \rightarrow \psi \\
\hline
\psi
\end{array}
\quad \text{MP}
\]

Fix a set \(\mathcal{F}\) of logical formulas. Let \(\mathcal{A} \subseteq \mathcal{F}\) be a predefined subset of formulas, which we will call \textit{logical axioms} of \(\mathcal{F}\). The axiomatic deductive system itself is the pair \((\mathcal{F}, \mathcal{A})\).

Given a proof of the deductive system, we split its premises into \textit{logical axioms} and \textit{non-logical axioms} depending on whether they belong to \(\mathcal{A}\) or not.

Note that we cannot have a complete axiomatic deductive system for first-order logic because of the eigenvariable condition in the rules \((\forall^+)\) and \((\exists^-)\).

Proposition 877. \textit{Given a deductive system, if} \(\Gamma \vdash \varphi\), \textit{then} \(\Gamma, \Delta \vdash \varphi\) \textit{for any formula} \(\varphi\) \textit{and any sets} \(\Gamma\) \textit{and} \(\Delta\).

\textit{If every formula in} \(\Delta\) \textit{is derivable from} \(\Gamma\), \textit{then the converse also holds:} \(\Gamma, \Delta \vdash \varphi\) \textit{implies} \(\Gamma \vdash \varphi\).

\textit{Proof.} If there exists a proof \(\varphi\) from \(\Gamma\), then adding additional axioms does not change anything.

The second part of the theorem has a tad more complicated proof. Assume that every formula in \(\Delta\) is derivable from \(\Gamma\) and that \(\Gamma, \Delta \vdash \varphi\).

For every \(\delta \in \Delta\), let \(P_\delta\) be a proof of \(\delta\) from members of \(\Gamma\) and let \(P_\varphi\) be a proof of \(\varphi\) from \(\Gamma \cup \Delta\).

Then, for every \(\delta \in \Delta\), we can replace the subtree of \(\delta\) with \(P_\delta\) to obtain a proof of \(\varphi\) from \(\Gamma\).

Therefore, \(\Gamma \vdash \varphi\). \qed

Definition 878. \textbf{The positive implicational propositional deductive system} is an extraordinarily simple \textit{axiomatic deductive system}.

It is based on the \textit{language of propositional logic}, but limited to formulas containing only the \textit{conditional connective} \(\rightarrow\), without any \textit{propositional constants} or \textit{negation}. Note that this is only a special case of \textit{positive formulas}.

The adjective “positive” in the name of the system refers to the impossibility to negate a formula. “Implicational” refers to the fact that all formulas are \textit{material implications} and the \textit{sole inference rule} is based on eliminating the connective.

The system has the following logical axiom schemas:

\textsuperscript{14}en: mode that by affirming affirms
(a) For every formula \( \varphi \), we can “introduce” an implication whose consequent is \( \varphi \) and whose antecedent is any other formula \( \psi \):

\[
\varphi \rightarrow (\psi \rightarrow \varphi). \tag{AX \rightarrow^+}
\]

(b) Implication is transitive:

\[
(\varphi \rightarrow (\psi \rightarrow \theta)) \rightarrow ((\varphi \rightarrow \psi) \rightarrow (\varphi \rightarrow \theta)). \tag{AX \rightarrow}
\]

**Example 879.** Fix any positive implicational formula \( \varphi \). We will construct a derivation of the implication

\[
\varphi \rightarrow \varphi. \tag{311}
\]

We derive the proof from the two logical axioms:

\[
\begin{align*}
(AX \rightarrow^+) & \quad (AX \rightarrow) \\
\vdots \quad \psi \mapsto (\varphi \rightarrow \varphi) & \quad \vdots \quad \psi \mapsto (\varphi \rightarrow \varphi) \\
\theta \mapsto \varphi & \quad (AX \rightarrow^+) \\
(313) & \quad (313) \rightarrow (((\varphi \rightarrow (\varphi \rightarrow \varphi)) \rightarrow (\varphi \rightarrow \varphi)) \\
(\varphi \rightarrow (\varphi \rightarrow \varphi)) \rightarrow (\varphi \rightarrow \varphi) & \quad \psi \mapsto \varphi \\
\varphi \rightarrow \varphi & \quad MP \\
\end{align*}
\]

where

\[
\varphi \rightarrow ((\varphi \rightarrow \varphi) \rightarrow \varphi). \tag{313}
\]

The only assumptions used in the derivation were logical axioms, hence (311) is a logical theorem.

**Definition 880.** We introduce two notions connecting derivability and satisfiability:

(a) If, for any closed formula \( \varphi \), derivability \( \vdash \varphi \) implies satisfiability \( \models \varphi \), we say that the deductive system is **sound** with respect to the semantical framework.

(b) Dually, if satisfiability \( \models \varphi \) implies derivability \( \vdash \varphi \) for any closed formula \( \varphi \), we say that the deductive system is **complete** with respect to the semantical framework.

We restrict our attention to closed formulas because we wish to avoid the problems described in remark 847. If we have a formula with free variables, we may simply take its universal closure.

**Proposition 881.** The positive implicational propositional deductive system is **sound** with respect to classical semantics.

**Theorem 882** (Syntactic deduction theorem). In the positive implicational deductive system, \( \Gamma, \psi \vdash \varphi \) holds if and only if \( \Gamma \vdash \psi \rightarrow \varphi \) holds.

This theorem also holds for propositional deductive systems which extend the positive implication system with compatible rules, as in the case of proposition 885 or definition 901.
Proof.

Proof of sufficiency. Suppose that $\Gamma, \psi \vdash \varphi$ and let $P$ be a proof of $\varphi$ from $\Gamma \cup \{\psi\}$. We will use theorem 792 (Structural induction on unambiguous grammars) on $\varphi$.

- First suppose that $\varphi$ is either a logical or a nonlogical axiom; in the latter case either $\varphi \in \Gamma$ or $\varphi = \psi$. In all three cases, the logical axiom $(AX \rightarrow^+)$ allows us to derive $\psi \rightarrow \varphi$ from $\Gamma$ using (MP).

- Otherwise, since the only rule is (MP), there exists some formula $\theta$ derivable from $\Gamma \cup \{\psi\}$ such that $P$ contains the formulas $\theta$ and for $\theta \rightarrow \varphi$. The inductive hypothesis holds for both, and hence $\Gamma \vdash \psi \rightarrow \theta$ and $\Gamma \vdash \psi \rightarrow (\theta \rightarrow \varphi)$. Let $P_1$ and $P_2$ be proofs corresponding to these two sequents.

We can now build the following proof of $\psi \rightarrow \varphi$ from $\Gamma$:

\[
\begin{array}{c}
(AX \rightarrow) \\
\vdots \\
P_2 \quad \left(\psi \rightarrow (\theta \rightarrow \varphi)\right) \rightarrow ((\psi \rightarrow \theta) \rightarrow (\psi \rightarrow \varphi)) \\
\hline
\left(\psi \rightarrow \theta\right) \rightarrow (\psi \rightarrow \varphi) \\
\hline
\psi \rightarrow \varphi
\end{array}
\]

Proof of necessity. Now suppose that $\Gamma \vdash \psi \rightarrow \varphi$. Then we can apply (MP) to obtain $\varphi$ from $\Gamma \cup \{\psi\}$. □

Definition 883. While the positive implicational propositional deductive system is simple, it is of more practical use to have all propositional connectives available. As it turns out, we cannot utilize complete families of Boolean functions unless we are dealing with classical semantics — see for example example 891 and example 894 for how proposition 815 (d) fails to hold.

Our goal is to define the (axiomatic) minimal propositional axiomatic deductive system, which would correspond to minimal logic. It is axiomatic in the sense that we do not use new rules to express the rest of the propositional syntax, but instead we need axiom schemas for each connective. The only exception is $\bot$, the axioms for which tend to change semantics by a lot — see theorem 887.

Axioms with $+$ in the superscript are called introduction axioms and axioms with $-$ are called elimination axioms.

The following axioms are essential in the sense that they cannot be defined in terms of others:

(a) The simplest axiom states that the constant $\top$ is itself an axiom:

\[
\top \quad (AX \top^+)
\]

[Aut20] def. 55.10
(b) Axioms for conjunction:

\[
\begin{align*}
\varphi \land \psi & \rightarrow \psi \quad \text{(AX} \land_L^-) \\
\varphi \land \psi & \rightarrow \varphi \quad \text{(AX} \land_R^-) \\
\varphi \rightarrow (\psi \rightarrow (\varphi \land \psi)) & \quad \text{(AX} \land^+) \\
\end{align*}
\]

(c) Axioms for disjunction:

\[
\begin{align*}
\varphi & \rightarrow (\varphi \lor \psi) \quad \text{(AX} \lor_L^+) \\
\psi & \rightarrow (\varphi \lor \psi) \quad \text{(AX} \lor_R^+) \\
(\varphi \rightarrow \theta) & \rightarrow ((\psi \rightarrow \theta) \rightarrow ((\varphi \lor \psi) \rightarrow \theta)) \quad \text{(AX} \lor^-) \\
\end{align*}
\]

The following axioms and are said to be “abbreviations” and do not affect semantics:

(d) The axioms for the biconditional is motivated via proposition 815 (e):

\[
\begin{align*}
(\varphi \rightarrow \psi) & \rightarrow (\psi \rightarrow \varphi) \rightarrow (\varphi \leftrightarrow \psi) \quad \text{(AX} \leftrightarrow^+) \\
(\varphi \leftrightarrow \psi) & \rightarrow (\varphi \rightarrow \psi) \quad \text{(AX} \leftrightarrow_L^-) \\
(\varphi \leftrightarrow \psi) & \rightarrow (\psi \rightarrow \varphi) \quad \text{(AX} \leftrightarrow_R^-) \\
\end{align*}
\]

(e) The axioms for negation is motivated via proposition 815 (a):

\[
\begin{align*}
\neg \varphi & \rightarrow (\varphi \rightarrow \bot) \quad \text{(AX} \neg^-) \\
(\varphi \rightarrow \bot) & \rightarrow \neg \varphi \quad \text{(AX} \neg^+) \\
\end{align*}
\]

Definition 884. Natural deductive systems are deductive systems whose set of rules allows discharging certain assumptions of the proof tree. These rules correspond to “bringing in” the sequent $\varphi \vdash \psi$ as a formula $\varphi \rightarrow \psi$, thus eliminating $\varphi$ as an assumption as justified by theorem 882 (Syntactic deduction theorem).

In a natural deduction system, all assumptions in a proof trees are labeled differently. When applying a rule that supports discharging assumptions, we add an additional label to the subproof that matches the label of the assumption which we discharge.

This allows us to distinguish between discharged assumptions and undischarged assumptions. We add a non-premise label to every discharged assumption so that it does not affect derivability.

Proposition 885. We define the minimal propositional natural deduction system, which is the natural deduction equivalent of the minimal propositional axiomatic deductive system.

(a) The following rules corresponds to the conditional axiom schemas in definition 878:
This rule is inspired by \((AX \to^+)\): 
\[
\begin{array}{c}
[\psi]^n \\
\vdots \\
n \frac{\varphi}{\psi \to^+ \varphi}
\end{array}
\]

This rule is merely a renaming of \((MP)\): 
\[
\frac{\varphi \to \psi}{\varphi \to^+ \psi} \quad \frac{\varphi}{\varphi \to^+}
\]

The additional notation in \((\to^+)\) means that the premise labeled with \(n\), if any, can be discharged.

Note that there is no rule corresponding to \((AX \to)\) because this axiom schema follows from \((\to^+\) and \((\to^-). Unlike in the axiomatic deductive system where \((AX \to)\) is used to prove theorem 882 (Syntactic deduction theorem), here we have a stronger connection between \(\to\) in the object language and \(\vdash\) in the metalanguage given by \((\to^+)\).

(b) The following rule corresponds to the axiom \((AX \top^+)\):
\[
\frac{\top}{\top^+}
\]

As discussed in definition 874 (b), applications of this rule have a non-premise label in order to prevent \(\top\) as an undischarged assumption.

(c) The following rules correspond to the conjunction axiom schemas in definition 883 (b):
\[
\frac{\varphi \quad \psi}{\varphi \land^+ \psi} 
\quad \frac{\varphi \land \psi}{\varphi \land^- \psi} 
\quad \frac{\varphi \land \psi}{\varphi \land^- \psi}
\]

(d) The following rules correspond to the disjunction axiom schemas in definition 883 (c):
\[
\frac{\varphi}{\varphi \lor \psi \lor^+} 
\quad \frac{\psi}{\varphi \lor \psi \lor^-} 
\quad \frac{[\varphi]^n \quad [\psi]^n}{\vdots \quad \vdots} 
\quad n \frac{\varphi \lor \psi \lor \theta \lor \theta}{\theta \lor^-}
\]

(e) The following rules correspond to the biconditional axiom schemas in definition 883 (d):
\[
\frac{[\varphi]^n \quad [\psi]^n}{\vdots \quad \vdots} 
\quad \frac{\varphi \leftrightarrow \psi \leftrightarrow^+ \psi}{\varphi \leftrightarrow \psi \leftrightarrow^- \psi} 
\quad \frac{\varphi \leftrightarrow \psi \leftrightarrow \varphi \leftrightarrow^- \varphi}{\varphi \leftrightarrow \psi \leftrightarrow \varphi \leftrightarrow^- \varphi}
\]

(f) The following rules correspond to the negation axiom schemas in definition 883 (e):
\[
\frac{[\varphi]^n}{\vdots} 
\quad \frac{\varphi}{\vdash \neg \varphi \neg^+} 
\quad \frac{\neg \varphi}{\vdash \neg \neg^-}
\]

Proof of correctness. We will prove that the axiomatic minimal propositional axiomatic deductive system is equivalent to the rules of natural deduction described in this proposition.
Proof of 885 (a). Consider first the axiom \((AX \rightarrow +)\). Fix two formulas \(\varphi\) and \(\psi\). Then \(\varphi \rightarrow (\psi \rightarrow \varphi)\) is an instance of \((AX \rightarrow +)\). Thus, we obtain \(\varphi \vdash \psi \rightarrow \varphi\) by applying (MP), which in turn shows the validity of the rule \((\rightarrow +)\).

The labeled assumption here is essential for showing that \((\rightarrow +)\) implies \((AX \rightarrow +)\). Without it we would have the rule
\[
\begin{align*}
\psi \\
\varphi \rightarrow \psi
\end{align*}
\]
which would not allow us to discharge the assumption \(\varphi\) when it is in fact immaterial for the validity of \(\psi\).

Now we will show that \((AX \rightarrow)\) can be derived using only the rules \((\rightarrow +)\) and \((\rightarrow -)\):

\[
\begin{array}{c}
[\varphi \rightarrow (\psi \rightarrow \theta)]^1 \\
\varphi \rightarrow \theta \\
\psi \\
[\varphi \rightarrow \theta]^2 \\
\rightarrow \\
[\varphi \rightarrow \psi]^3 \\
\varphi \\
\rightarrow \\
[\varphi]^2 \\
\rightarrow
\end{array}
\]

\[
\begin{array}{c}
\theta \\
\varphi \rightarrow \theta \\
[\varphi \rightarrow \theta]^3 \\
\rightarrow \\
(AX \rightarrow) \\
[\varphi \rightarrow \theta] \\
\rightarrow \\
\rightarrow
\end{array}
\]

Proof of 885 (b). Obvious.

Proof of 885 (c). The rule \((AX ^+ \land)\) is equivalent by more readable than proving \(\{\varphi, \psi\} \vdash \varphi \land \psi\) directly. Indeed, compare it to
\[
\begin{align*}
\varphi & \quad \text{(AX ^+ \land)} \\
& \quad \text{MP} \\
\varphi \rightarrow (\varphi \land \psi) \\
& \quad \text{MP} \\
\varphi \land \psi
\end{align*}
\]
which is a derivation of \(\varphi \land \psi\) from \(\{\varphi, \psi\}\) using the axiomatic system. The other direction is also simple:

\[
\begin{array}{c}
[\varphi]^1 \\
\varphi \land \psi \\
\rightarrow \\
[\psi]^2 \\
\varphi \land \psi \\
\rightarrow \\
\psi \rightarrow (\varphi \land \psi) \\
(AX ^+ \land) \\
\rightarrow \\
\rightarrow
\end{array}
\]

The other two rules are trivially connected to the corresponding axioms using a single application of (MP).

Proof of 885 (d). For a more complicated example, consider \((AX \lor -)\). We have
\[
\begin{align*}
(AX \lor -) & \quad \varphi \rightarrow \theta \\
(\psi \rightarrow \theta) \rightarrow ((\varphi \lor \psi) \rightarrow \theta) & \quad \text{MP} \\
(\varphi \lor \psi) \rightarrow \theta \\
\theta & \quad \text{MP}
\end{align*}
\]

445
The assumptions of this derivation are $\varphi \rightarrow \theta$, $\psi \rightarrow \theta$ and $\varphi \lor \psi$. Instead of adding them directly as premises of the inference rule ($\lor^{-}$), we replace the conditional $\rightarrow$ with labeled assumptions that correspond to $\varphi \vdash \theta$ and $\psi \vdash \theta$.

We can prove that ($\lor^{-}$) implies ($AX \lor^{-}$) analogously to (315).

The other two rules are again trivial to obtain from the corresponding axioms and vice versa.

**Proof of 885 (e).** Analogous to what we have already shown.

**Proof of 885 (f).** ($\neg^+$) is obtained from ($AX \neg^+$) by applying (MP) once and ($\neg^-$) is obtained from ($AX \neg^+$) by applying (MP) twice. Using the rules to derive the axioms is similar to (315).

**Proposition 886.** In deductive systems that extend the minimal propositional natural deduction system, we have $\psi_1, \psi_2 \vdash \varphi$ if and only if $(\psi_1 \land \psi_2) \vdash \varphi$.

**Proof.**

**Proof of sufficiency.** If $\psi_1, \psi_2 \vdash \varphi$, then

$$
\frac{\psi_1 \land \psi_2}{\psi_1} (\land^{-} R) \quad \frac{\psi_1 \land \psi_2}{\psi_2} (\land^{-} R)
$$

\[ \vdots \]

$$
\varphi
$$

**Proof of necessity.** If $(\psi_1 \land \psi_2) \vdash \varphi$, then

$$
\frac{\psi_1 \land \psi_2}{\psi_1} (\land^{+}) \quad \frac{\psi_1 \land \psi_2}{\psi_2} (\land^{+})
$$

\[ \vdots \]

$$
\varphi
$$

**Theorem 887.** Consider the following propositional formula schemas:

(a) **Double negation elimination:**

$$
\neg \neg \varphi \rightarrow \varphi.
$$

The semantic counterpart to this law is **proposition 815 (b).**

(b) **Ex falso quodlibet**, also known as the **principle of explosion**:

$$
\bot \rightarrow \varphi
$$

(c) **Pierce’s law**:

$$
((\varphi \rightarrow \psi) \rightarrow \varphi) \rightarrow \varphi
$$

\[^{15}\text{en: from falsity everything follows} \]

446
(d) The **law of the excluded middle**:
\[ \varphi \lor \neg \varphi \]  
(AX LEM)

(e) The **law of non-contradiction**:
\[ \neg (\varphi \land \neg \varphi) \]  
(AX LNC)

Assuming the minimal propositional axiomatic deductive system, we have the following derivations:

\[
\begin{array}{c}
\text{(AX DNE)} \\
\text{(AX Pierce)} \quad \text{(AX EFQ)} \\
\text{(AX LEM)} \quad \text{(AX LNC)}
\end{array}
\]

As it turns out, (AX LNC), which is often associated with intuitionistic logic, is a theorem of **minimal logic**.
Conversely, (AX EFQ) and (AX LEM) together can be used to derive (AX DNE).

**Proof.** Most proofs are given in [DM16, prop. 3] and [DM16, prop. 13]. We will only show that (AX LNC) is strictly weaker than (AX EFQ).

For any formula \( \varphi \), we have the natural deduction proof that (AX LNC) is a tautology:

\[
\begin{array}{c}
\frac{\varphi \land \neg \varphi}{\varphi} \quad \frac{\varphi \land \neg \varphi}{\neg \varphi} \\
\[1\] \quad \[2\] \\
\frac{1}{\bot} \quad \frac{\neg \varphi}{\neg \neg \neg} \\
\frac{\bot}{\neg (\varphi \land \neg \varphi)}
\end{array}
\]

Hence, (AX LNC) is a theorem of **minimal logic**. If it were to imply (AX EFQ), then minimal and intuitionistic logic would be equivalent, which would contradict [DM16, prop. 3]. Therefore, (AX LNC) is indeed strictly weaker than (AX EFQ).

**Proposition 888.** In the minimal propositional natural deduction system, we have
\[ \varphi \rightarrow \psi \vdash \neg \psi \rightarrow \neg \varphi \]  
(316)  
(AX DNE), \[ \neg \varphi \rightarrow \neg \psi \vdash \psi \rightarrow \varphi \]  
(317)

**Proof.** We will only prove (316). The derivability (317) can be proved in the same way except that we would use \((\neg^+)\) rather than (DNE).

\[
\begin{array}{c}
\frac{\varphi \rightarrow \psi}{\psi} \quad \frac{\varphi}{[-\psi]} \\
\[1\] \quad \[2\] \\
\frac{\neg \varphi}{\neg \neg \neg} \quad \frac{\neg \neg \neg}{\neg \neg \neg} \\
\frac{\bot}{\neg \psi} \quad \frac{\bot}{\neg \neg \neg} \\
\frac{\bot}{\neg \neg \neg} \quad \frac{\bot}{\neg \neg \neg}
\end{array}
\]

\[\square\]
**Definition 889.** The intutionistic propositional natural deduction system extends the minimal propositional natural deduction system with the rule

\[ \frac{\bot}{\varphi} \text{ EFQ} \]

This corresponds to the axiom (AX EFQ), which we can add to the minimal propositional axiomatic deductive system.

The corresponding semantics are defined in definition 890 and their link with the deductive system is given in theorem 892.

**Definition 890.** We define Heyting semantics for propositional formulas similarly to how it is done with classical Boolean semantics in definition 813, except that instead of using a Boolean algebra we use a more general Heyting algebra.

Logical negations depend on complements in Boolean algebras. Since Heyting algebras do not have complements, we instead use pseudocomplements.

Fix a Heyting algebra \( \mathcal{K} = (H, \sup, \inf, T, F, \rightarrow) \). Propositional interpretations in Heyting semantics may take any value in \( X \), as can formula valuations.

Given an interpretation \( I \) and a formula \( \varphi \), we define \( \varphi[I] \) via (256), the sole difference being that negation valuation is defined via the pseudocomplement:

\[ (\neg \psi)[I] := \bar{\psi}[I]. \]

We say that \( I \) satisfies \( \varphi \) if \( \varphi[I] = T \). Thus, if the valuation of \( \varphi \) takes any value in \( H \setminus \{T\} \), then \( I \) does not satisfy \( \varphi \), but that does not necessarily mean that \( I \) satisfies \( \neg \varphi \).

Then \( \Gamma \) entails \( \varphi \) if, for every \( \psi \in \Gamma \) and every interpretation \( I \) in every Heyting algebra, we have \( \varphi[I] = \psi[I] \).

It is important that different Heyting algebras may provide different semantics — see example 891 for an example of what is impossible in a Boolean algebra.

**Example 891.** Let \( \mathcal{X} \) be an extension of the trivial Boolean algebra \( \{T, F\} \) with the “indeterminate” symbol \( N \). That is, the domain of \( \mathcal{X} \) is \( \{F, N, T\} \) and the order is \( F \leq N \leq T \).

The pseudocomplement of \( N \) is

\[ \bar{N} \overset{(519)}{=} \sup\{a \in X \mid a \land N = \bot\} = F. \]

Consider any propositional interpretation \( I \) such that \( I(P) = N \).

Then the valuation of \( \text{(AX LEM)} \) is

\[ (P \lor \neg P)[I] = \sup\{P[I], \bar{P}[I]\} = \sup\{N, \bar{N}\} = \sup\{N, F\} = N. \]

Therefore, \( \text{(AX LEM)} \) does not hold.

**Theorem 892.** The intutionistic propositional deductive system is sound and complete with respect to Heyting semantics. To elaborate,

(a) If \( \models \varphi \), then \( \models \varphi \) for every Heyting algebra.
(b) If $\models \varphi$ in every Heyting algebra, then $\vdash \varphi$.

**Definition 893.** Since arbitrary Heyting algebras can be cumbersome to come up with when used for propositional Heyting semantics, we can instead utilize example 1271 and define **topological semantics** for some nonempty topological space.

The truth values of interpretations and valuations are then open sets in some topological space and a formula is said to be valid if its valuation is the whole space.

**Example 894.** Let $U$ be an open set in the standard topology in $\mathbb{R}$. We will examine (AX LEM) with respect to topological semantics for $\mathbb{R}$. Due to example 1271, given any propositional interpretation $I$ such that $I(\varphi) = U$, we have

$$(\varphi \lor \neg\varphi) [I] = P[I] \cup P[\neg I] = U \cup \overline{U} = U \cup \text{int}(\mathbb{R} \setminus U).$$

If $U = \emptyset$, then $(\varphi \lor \neg\varphi)[I] = \mathbb{R}$ and (AX LEM) holds. If $U = (0, 1)$, then $(\varphi \lor \neg\varphi)[I] = \mathbb{R} \setminus \{0, 1\}$ and (AX LEM) does not hold.

Compare this result with example 891.

**Definition 895.** Another semantics for the intuitionistic propositional deductive system is the **Brouwer-Heyting-Kolmogorov interpretation**.

It uses a less formal approach than Heyting algebra semantics that is based on the notion of a “construction”, which is also why it is sometimes called **constructive logic**.

(a) We assume that we know what constitutes a construction of propositional variables.

(b) There is no construction of $\bot$ and no construction of $\top$ is needed.

(c) A construction of $\psi_1 \lor \psi_2$ is a pair $(k, M)$, where $k = 1, 2$ and $M$ is a construction of $\psi_m$ if and only if $k = m$. The notion of a pair here is informal.

(d) A construction of $\psi_1 \land \psi_2$ is a pair $(M_1, M_2)$, where $M_k$ is a construction of $\psi_k$ for $k = 1, 2$.

(e) A construction of $\psi_1 \rightarrow \psi_2$ is a function that converts a construction of $\psi_1$ into a construction of $\psi_2$. The notion of a function here is informal.

The negation $\neg \psi$ that corresponds to pseudocomplements in Heyting algebra semantics corresponds to the metastatement “a construction of $\psi$ is impossible” under the Heyting-Brouwer-Kolmogorov interpretation.

If the set $\Gamma$ of formulas does not derive $\varphi$, we say that $\varphi$ is non-constructive under the axioms $\Gamma$.

**Remark 896.** Since the Heyting-Brouwer-Kolmogorov interpretation is not very formal, we cannot properly prove its, soundness or completeness with respect to the intuitionistic propositional deductive system.

Nevertheless, we generally accept the interpretation and conflate “constructive” and “intuitionistic” statements.
Example 897. Theorem 1032 (Zermelo’s well-ordering theorem) in ZFC does not provide a way to well-order an arbitrary set. The theorem relies on the axiom of choice, whose consequence theorem 991 (Diaconescu-Goodman-Myhill theorem) implies the law of the excluded middle (LEM) assuming the nonlogical axioms of ZFC.

Since LEM may not hold in intuitionistic logic, it follows that both theorem 1032 (Zermelo’s well-ordering theorem) and the axiom of choice itself should not in general hold under the Heyting-Brouwer-Kolmogorov interpretation, hence by the terminology in definition 895, theorem 1032 (Zermelo’s well-ordering theorem) is a non-constructive theorem.

Definition 898. In order to obtain a deductive system that matches classical propositional semantics, we may extend the minimal propositional natural deduction system with the rule

\[ \vdash \neg \phi \]

\[ \vdots \]

\[ n \vdash \phi \text{ DNE} \]

This corresponds to the axiom (AX DNE), which we can add to the minimal propositional axiomatic deductive system. As per theorem 887, we can instead add (AX LEM) to the intuitionistic propositional axiomatic deductive system, since

\[ (AX \text{ LEM}), (AX \text{ EFQ}) \vdash (AX \text{ DNE}) \]

We call this, very simply, the (classical) propositional deductive system.

Theorem 899 (Glivenko’s double negation theorem). A formula \( \phi \) is derivable in the classical propositional natural deduction system if and only if its double negation \( \neg \neg \phi \) is derivable in the intuitionistic propositional natural deduction system.

Theorem 900. The classical propositional natural deduction system is sound and complete with respect to classical semantics.

Definition 901. If we wish to work with first-order logic rather than merely propositional logic, we must extend the classical propositional natural deduction system. We call this, very simply, the (classical) first-order natural deduction system.

(a) We first add the following two inference rules for quantification:

\[ \frac{\phi}{\forall \xi . \phi} \]

\[ \frac{[\phi]^n \vdash \psi}{n \vdash \exists \xi . \phi \text{ DNE}} \]

Here \( \xi \) is a variable that is not free in any undischarged assumption in the proof of \( \phi \) (it may be free in \( \phi \) as long as \( \phi \) is not itself an undischarged assumption). A variable \( \xi \) satisfying these conditions is called an eigenvariable of the rule.

These rules are the primary motivation for inference rules accepting proof trees rather than only formulas — see definition 871 (b) and definition 874 (b). See example 902 (b) for why this condition is important.
(b) We add two inference rules, where \( \tau \) is an arbitrary term:

\[
\begin{align*}
\forall \xi. \varphi & \quad \varphi[\xi \mapsto \tau] \quad \varphi[\xi \mapsto \tau] \\
\end{align*}
\]

\( \exists \xi. \varphi \)

Compare this to proposition 844.

(c) Finally, we also add three rules for formal equality:

\[
\begin{align*}
\tau \equiv \tau & \quad \equiv^+ \\
\tau \equiv \sigma & \quad \varphi[\xi \mapsto \tau] \quad \varphi[\xi \mapsto \sigma] \\
\varphi[\xi \mapsto \tau] \quad \varphi[\xi \mapsto \sigma] & \quad \equiv^- \\
\end{align*}
\]

**Example 902.**

(a) We explicitly forbid the syntactic equivalent of proposition 843 in order to avoid invalid proofs like example 902 (b). Consider the proof

\[
\begin{align*}
&\varphi[1] \\
&\forall \xi. \varphi \\
\end{align*}
\]

The problem here is that \( \varphi \) is itself an undischarged assumption, hence \( (\forall^+) \) is actually inapplicable here, and the proof is invalid.

(b) To see why the eigenvariable conditions in definition 901 (a) are essential, consider the following proof of \( \forall \xi. \varphi \) from \( \exists \xi. \varphi \):

\[
\begin{align*}
&\exists \xi. \varphi[1] \\
&\forall \xi. \varphi \quad \forall^+ \\
&\varphi \quad \exists^- \\
\end{align*}
\]

This proof relies on example 902 (a), which we have already demonstrated to be invalid.

(c) Another invalid proof, in case \( \xi \in \text{Free}(\varphi) \), is

\[
\begin{align*}
&\exists \xi. \varphi[1] \\
&\varphi \quad \exists^- \\
\end{align*}
\]

(d) On the other hand, \( \exists \xi. \varphi \) can easily be derived from \( \forall \xi. \varphi \):

\[
\begin{align*}
&\forall \xi. \varphi \\
&\varphi \quad \exists^+ \\
\end{align*}
\]

(e) It is also valid to perform the completely meaningless derivation:

\[
\begin{align*}
&\forall \xi. \varphi \\
&\varphi \quad \forall^+ \\
\end{align*}
\]

451
**Proposition 903.** For any formula \( \varphi \) and any variable \( \xi \) over \( \mathcal{L} \), we have the following inter-derivable pairs:

\[
\neg \forall \xi . \varphi \text{ and } \exists \xi . \neg \varphi \tag{318}
\]

\[
\neg \exists \xi . \varphi \text{ and } \forall \xi . \neg \varphi \tag{319}
\]

**Proof.** We will only show (318). First,

\[
\begin{array}{c}
\neg \forall \xi . \varphi \\
\text{[\( \forall \xi . \varphi \)]}^1 \\
\text{(DNE)} \\
\text{[\( \neg \varphi \)]}^2 \\
\text{[\( \neg \neg \varphi \)]} \\
\text{(\( \exists \neg \neg \varphi \))}
\end{array}
\]

Conversely,

\[
\begin{array}{c}
\forall \xi . \varphi \\
\text{[\( \forall \neg \varphi \)]}^1 \\
\text{(\( \neg \neg \varphi \))} \\
\text{[\( \neg \neg \neg \varphi \)]} \\
\text{(\( \exists \neg \neg \neg \varphi \))}
\end{array}
\]

\[\square\]

**Theorem 904.** The classical first-order natural deduction system is sound and complete with respect to classical semantics.

The completeness part is known as “Gödel’s completeness theorem” and requires an elaborate proof.
12.8. Logical theories

**Definition 905.** The closure of the set $\Gamma$ of formulas in the first-order language $\mathcal{L}$ is the set

$$\text{cl}(\Gamma) := \{ \varphi \in \text{Form} \mid \Gamma \vDash \varphi \}.$$

Due to remark 833, this definition also holds for formulas in propositional logic. The set $\Gamma$ is called **closed** if it equals its own closure. A closed set of closed formulas is also sometimes called a **theory** in e.g. [Aut20, def. 33.1], but we will not put this restriction because it is not clear what are the axioms of an arbitrary closed set of formulas. Instead, we will use the term “theory” to mean the set of axioms itself.

As discussed in remark 847, we are only interested in closed formulas. If a formula is not closed, we instead consider its universal closure.

If it is necessary to distinguish between derivability and entailment, we can use the syntactic closure $\text{cl} \vdash$ and semantic closure $\text{cl} \vDash$.

(a) We say that the set $\Gamma$ is **axiomatized** by $\Delta$ if $\Gamma = \text{cl}(\Delta)$.

(b) The set $\Gamma$ of formulas is said to be **complete** if every for every formula in $\varphi$, either $\Gamma \vDash \varphi$ or $\Gamma \vDash \neg \varphi$.

This is not to be confused with definition 880 (b), which defines completeness of a deductive system via how it relates to semantics.

**Definition 906.** Given a deductive system, the set $\Gamma$ of formulas is said to be **inconsistent**\(^{16}\) if $\Gamma \vdash \bot$ and **consistent** otherwise.

**Proposition 907.** Assuming classical logic, a theory $\Gamma$ is unsatisfiable if and only if it is semantically inconsistent.

**Proof.**

**Proof of sufficiency.** Assume first that $\Gamma$ is unsatisfiable. Then for all zero models of $\Gamma$ any formula is satisfied vacuously, in particular that any model of $\Gamma$ satisfies $\bot$. Thus, $\Gamma \vDash \bot$ and, by theorem 904, $\Gamma \vdash \bot$. Therefore, $\Gamma$ is inconsistent.

**Proof of necessity.** Let $\Gamma$ be inconsistent and suppose that $\mathcal{X} = (X, I)$ is a model of $\Gamma$. Fix any valuation $v$ in $\mathcal{X}$. Since $\Gamma \vDash \bot$, theorem 904 implies that $\bot \llbracket v \rrbracket = T$.

By the definition of formula valuation, however, we have $\bot \llbracket v \rrbracket = F$. The obtained contradiction shows that $\mathcal{X}$ cannot be a model of $\Gamma$ and since this structure was chosen arbitrarily, we conclude that $\Gamma$ is unsatisfiable. \hfill \Box

**Theorem 908** (First-order compactness theorem). Assuming classical logic, for a theory $\Gamma$ we have $\Gamma \vDash \varphi$ for the closed formula $\varphi$ if and only if there exists a finite subset of $\Delta \subseteq \Gamma$ such that $\Delta \vDash \varphi$.

**Proof.**

\(^{16}\) bg: противоречива, ru: противоречивая
Proof of sufficiency. Suppose that $\Gamma \models \varphi$. By theorem 904, $\Gamma \vdash \varphi$. Every proof tree is finite, and thus $\varphi$ can be derived from only the finitely many premises $\Delta$ of the proof. Then $\Delta \not\vdash \varphi$ and, by the other direction of theorem 904, we have $\Delta \not\models \varphi$.

Proof of necessity. If $\Delta \models \varphi$ for some finite $\Delta \subseteq \Gamma$, then necessarily $\Gamma \models \varphi$.

Corollary 909. If every finite subset of $\Gamma$ is satisfiable, then $\Gamma$ itself is satisfiable.

This is also called “the compactness theorem”.

Proof. We can restate the theorem using its contraposition: If $\Gamma$ is unsatisfiable, then there exists a finite subset $\Delta \subseteq \Gamma$ that is also unsatisfiable. Due to proposition 907, this can be also restated as: If $\Gamma$ is inconsistent, then there exists a finite subset $\Delta \subseteq \Gamma$ that is also inconsistent.

But the last formulation is a special case of theorem 908 (First-order compactness theorem) with $\varphi = \bot$.

Theorem 910 (Downward Löwenheim-Skolem theorem). Let $\Gamma$ be a first-order theory over some arbitrary language. Suppose that $\Gamma$ has a model of infinite cardinality.

Then $\Gamma$ also has a countable model.

Example 911 (Skolem’s paradox). From theorem 910 (Downward Löwenheim-Skolem theorem) it follows that ZFC, if it is consistent, has a model at is at most countably infinite. The axiom of infinity states that no model of ZFC is finite. Therefore, there exists a model of ZFC that is countably infinite. But we can construct uncountable sets in ZFC.

Therefore, either:

- Uncountable sets within the metatheory are fundamentally different from those within the object theory.
- Uncountable sets are paradoxical and ZFC must disallow them in order to be consistent.
- ZFC is inconsistent even when restricted to countable sets.

Theorem 912 (Upward Löwenheim-Skolem theorem). Let $\Gamma$ be a first-order theory over some arbitrary language. Suppose that $\Gamma$ has a model of infinite cardinality $\kappa$.

Then for every infinite cardinal $\mu \geq \text{card}(\Gamma)$, the theory $\Gamma$ also has a model of cardinality $\mu$.

Definition 913. Let $\mathcal{L}$ be a first-order language and let $\Gamma$ be a nonempty set of formulas. We describe the categories of small models for $\Gamma$ as the concrete category. See remark 914 for why it is a reasonable restriction for $\Gamma$ to be a set of positive formulas.

Suppose that we are given a Grothendieck universe $\mathcal{U}$, which is safe to assume to be the smallest suitable one as explained in definition 1112.

- The set of objects is the set of all $\mathcal{U}$-small models of $\Gamma$.
- The set of morphisms between two models is the set of all structure homomorphisms between them.
- Composition of morphisms is the usual function composition.
The identity morphism on a model is the identity function on its universe.

**Remark 914.** We usually want \( \Gamma \) to be a set of positive formulas because, in general, homomorphisms are not injective and the homomorphic image of a model may fail to be a model. **Proposition 863** shows that if all formulas in \( \Gamma \) are positive, the image of any homomorphism is again a model and thus an object in the category.

**Example 915.** This is an incomplete list of categories of small models of first-order theories that can be found in this document:

- The categories \( \text{Set}_* \), \( \text{Inv} \), \( \text{Mag} \), \( \text{Mag}_* \), \( \text{Mon} \), \( \text{Grp} \) and \( \text{Ab} \), whose relations are crucial for the definition and properties of groups.
- The category \( \text{SRing} \) of semirings, \( \text{SMod}_R \) of semimodules, \( \text{Ring} \) of rings and \( \text{Mod}_R \) of modules, which are all based on commutative monoids.
- The categories \( \text{Pos} \) in order theory and the related \( \text{PreOrd} \), \( \text{Tos} \), \( \text{Lat} \), \( \text{Heyt} \) and \( \text{Bool} \) in lattice theory.

In contrast:

- We define the category \( \text{Top} \) of topological spaces and all of its related categories within set theory without a corresponding first-order theory.
- The category \( \text{Set} \) of sets with respect to either ZFC, ZFC+U or naïve set theory is not the same as the category of \( \mathcal{U} \)-small models of set theory. Instead, it is a category within a fixed set theory. Within the metatheory of this document, we work within a fixed model of ZFC+U with respect to classical logic.
- Similarly, we do not care about models of Peano arithmetic enough to study its category of \( \mathcal{U} \)-small models. Instead, we only use a single model and denote it by \( \mathbb{N} \).

**Definition 916.** For an object \( X \) in some category, define an equivalence relation on monomorphisms \( f : A \to X \) and \( g : B \to X \) with codomain \( X \) to hold if \( f \) factors through \( g \) and vice versa. We call the equivalence classes of this relation subobjects of \( X \).

Dually, for the relation where epimorphisms \( f : X \to A \) and \( g : X \to B \) factors through each other, we say that the equivalence classes are quotient objects of \( X \).

**Proposition 917.** Fix a category of small models over a first-order theory.

(a) Every nonempty structure embedding (injective homomorphism) is a categorical monomorphism.

In particular, via the canonical inclusion, every submodel of \( X \) is a categorical subobject of \( X \).

A partial converse always holds — every split monomorphism is injective.

If the forgetful functor \( U : C \to \text{Set} \) has a left adjoint, as it often does, then every monomorphism is injective. Otherwise, a homomorphism may be left invertible with respect to homomorphisms but not general functions.
(b) **Dually, every structure projection** (surjective homomorphism) is a categorical epimorphism.

In particular, via the canonical inclusion, every surjective function of $\mathcal{X}$ is a categorical quotient object of $\mathcal{X}$.

A partial converse always holds — every split epimorphism is surjective.

**Proof.** All follow from proposition 1199. □

**Definition 918.** Assume some fixed deductive system in propositional or first-order logic. Let $\Gamma$ be a closed set of closed formulas within the corresponding logic. Then $(\Gamma, \vdash)$ is a preordered set. The **Lindenbaum-Tarski algebra** of the theory $\Gamma$ is the partially ordered set obtained from $(\Gamma, \vdash)$ using proposition 1226.

More concretely, the Lindenbaum-Tarski algebra of $\Gamma$ is a quotient set of $\Gamma$ by the relation interderivability and endowed with the partial order

$$[\varphi] \leq [\psi] \text{ if and only if } \varphi \vdash \psi. \quad (320)$$

Of course, we can define the algebra using entailment rather than derivability, but in the cases we consider, the two are equivalent and derivability is simpler to work with.

**Proof.** The correctness of (320), i.e. the fact that the relation $\leq$ does not depend on the choice of representatives from the quotient sets, follows from proposition 1226.

We must only demonstrate that $(\Gamma, \vdash)$ is indeed a preordered set. Reflexivity of $\vdash$ follows from definition 876 and transitivity follows from proposition 877. □

**Proposition 919.** Assume that we are working in the intuitionistic propositional natural deduction system. The **Lindenbaum-Tarski algebra** of a closed set of closed formulas $\Gamma$ then is a Heyting algebra.

In the classical derivation system, the algebra is instead a Boolean algebra.

In the minimal derivation system, we have an unbounded lattice with only a top element, but no bottom. Consequently, conditionals and pseudocomplements may fail to exist.

Explicitly:

(a) The **join** of the equivalence classes $[\varphi_1]$ and $[\varphi_2]$ is the class $[\varphi_1 \lor \varphi_2]$ of their disjunction.

(b) The class of **contradictions** $[\bot]$ is the bottom element.

(c) Similarly to joins, the **meet** of $[\varphi_1]$ and $[\varphi_2]$ is the equivalence class $[\varphi_1 \land \varphi_2]$ of their conjunction.

(d) The class of **tautologies** $[\top]$ is the top element.

(e) The **conditional** of $[\varphi_1]$ and $[\varphi_2]$ is the equivalence class $[\varphi_1 \rightarrow \varphi_2]$. 

456
(f) The pseudocomplement $\overline{\psi}$ of $[\psi]$ equals $\lnot \psi$.

In the classical derivation system this pseudocomplement is a complement, i.e. it satisfies (520) and (521).

Proof:

**Proof of 919 (a).** We will show that $[\psi_1 \lor \psi_2]$ is the supremum of $[\psi_1]$ and $[\psi_2]$.

The inference rule $\left(\lor^+\right)$ implies that $\psi_1 \vdash \psi_1 \lor \psi_2$ and $\left(\lor^-\right)$ implies that $\psi_2 \vdash \psi_1 \lor \psi_2$. Thus, $\psi_1 \lor \psi_2$ is an upper bound for both $\psi_1$ and $\psi_2$ under the ordering $\vdash$.

Let $\varphi$ be any formula in $\Gamma$ such that $\psi_1 \vdash \varphi$, $\psi_2 \vdash \varphi$ and $\varphi \vdash (\psi_1 \lor \psi_2)$. Then the following instance of $\left(\lor^-\right)$

\[
\begin{array}{ccc}
[\psi_1]^1 & [\psi_2]^1 & \vdash \\
\vdots & \vdots & \\
[\psi_1 \lor \psi_2] & \varphi & \varphi \\
\end{array}
\]

demonstrates that $\psi_1, \psi_2, (\psi_1 \lor \psi_2) \vdash \varphi$. Hence, from proposition 877 it follows that $(\psi_1 \lor \psi_2) \vdash \varphi$.

Obviously interderivability is not affected by the choice of representatives from $[\psi_1]$ and $[\psi_2]$, hence $[\varphi] = [\psi_1 \lor \psi_2]$ and this is indeed the supremum of $[\psi_1]$ and $[\psi_2]$.

**Proof of 919 (d).** The inference rule $\left(\top^+\right)$ shows that $[\varphi] \leq [\top]$ for any formula $\varphi$ and $[\top]$ is the top element.

**Proof of 919 (c).** Let $\psi_1$ and $\psi_2$ be any formulas in $\Gamma$. The inference rule $\left(\land^-\right)$ implies that $\psi_1 \land \psi_2 \vdash \psi_2$ and $\left(\land^-\right)$ implies that $\psi_1 \land \psi_2 \vdash \psi_1$. Thus, $\psi_1 \land \psi_2$ is a lower bound for both $\psi_1$ and $\psi_2$ under the ordering $\vdash$.

We must show that $\psi_1 \land \psi_2$ is interderivable with the greatest lower bound. Let $\varphi$ be a formula in $\Gamma$ such that $\varphi \vdash \psi_1$, $\varphi \vdash \psi_2$ and $(\psi_1 \land \psi_2) \vdash \varphi$. We will show that $\varphi \vdash (\psi_1 \land \psi_2)$.

The rule $\left(\land^+\right)$ implies that

$\psi_1, \psi_2 \vdash \psi_1 \land \psi_2$.

But $\varphi$ derives both $\psi_1$ and $\psi_2$, hence due to proposition 877,

$\varphi \vdash (\psi_1 \land \psi_2)$.

Analogously to the proof of proposition 919 (a), we conclude that the choice of representatives of the equivalence classes is irrelevant and that $[\varphi] = [\psi_1 \land \psi_2]$ is the infimum of $[\psi_1]$ and $[\psi_2]$.

**Proof of 919 (b).** That $[\bot]$ is the bottom is a restatement of (AX EFQ).

**Proof of 919 (e).** Restating (518), we must prove that $[\psi_1 \rightarrow \psi_2]$ equals

\[
([\psi_1] \rightarrow [\psi_2]) = \sup\{[\varphi] \mid \varphi \in \Gamma \text{ and } (\varphi \land \psi_1) \vdash \psi_2\}.
\]

An equivalent condition for $\varphi$ to be in $\Phi$ is, due to proposition 886,

$\varphi, \psi_1 \vdash \psi_2$. 

457
Then by the deduction theorem,

$$\varphi \vdash (\psi_1 \rightarrow \psi_2).$$

Thus, $$[\psi_1 \rightarrow \psi_2]$$ is an upper bound of $$\Psi$$.

It remains to show that $$(\psi_1 \rightarrow \psi_2) \in \Psi$$. Both $$\psi_1$$ and $$(\psi_1 \rightarrow \psi_2)$$ follow from the formula $$((\psi_1 \rightarrow \psi_2) \land \psi_1)$$ and by applying (MP), we obtain

$$\psi_1, (\psi_1 \rightarrow \psi_2) \vdash \psi_2.$$

From proposition 886,

$$((\psi_1 \rightarrow \psi_2) \land \psi_1) \vdash \psi_2.$$

Therefore, $$[\psi_1 \rightarrow \psi_2] \in \Phi$$ and it is indeed the supremum of $$\Psi$$.

**Proof of 919 (f).** The pseudocomplement $$\widehat{\psi}$$ is, by definition,

$$[\widehat{\psi}] = [\psi] \rightarrow [\bot].$$

From what we have already proved, we can conclude that $$[\widehat{\psi}] = [\psi \rightarrow \bot]$$. From definition 883 (e) it follows that the formulas $$\psi \rightarrow \bot$$ and $$\neg \psi$$ are interderivable, thus $$[\widehat{\psi}] = [\neg \psi]$$.

If we are working in classical logic where (AX LEM) holds, then

$$\sup\{[\psi], [\widehat{\psi}]\} \overset{919 (a)}{=} [\psi \lor \neg \psi] \overset{(AX LEM)}{=} [\top],$$

which proves (520).

The dual law (521)

$$\inf\{[\psi], [\widehat{\psi}]\} \overset{919 (c)}{=} [\psi \land \neg \psi] \overset{(AX LNC)}{=} [\neg \top] = [\widehat{\top}] \overset{(519)}{=} [\bot].$$

$$\square$$

**Remark 920.** Notice that the proof of proposition 919 relies entirely on the derivation system. On the other hand, we define Heyting semantics for intuitionistic formulas.

Thus, we have shown that Heyting algebras arise naturally from the intuitionistic natural deduction system and that their role as a semantic framework is completely justified.
13. Set theory

Sets are ubiquitous in mathematics yet set theory itself is quite complicated. Attention needs to be put to define a logical theory of sets that is both useful and consistent.

We first use the simplicity of naïve set theory to introduce some fundamental definitions. This theory turns out to be inconsistent due to theorem 932 (Russell’s paradox).

We later introduce the more, sophisticated Zermelo-Fraenkel set theory (ZFC). It is unclear whether the latter theory is consistent, however no contradictions have yet been discovered. Furthermore, the restrictions of ZFC (mostly the axiom of foundation) actually make it possible to define a model of Peano arithmetic and extend the concept of a number to ordinals and cardinals. We then discuss von Neumann’s cumulative hierarchy, which can be used to build models of set theory.

We also extend ZFC with Grothendieck universes to obtain the theory ZFC+U, upon which we construct the concept of a category.

By “set theory” we mean naïve set theory, ZF, ZFC, ZFC+U or further variations.

Remark 921. The relation between first-order logic and set theory is remarkably circular.

• Set theory is defined as a theory of first-order logic.

• First-order logic itself is defined via sets, for example via the language of first-order logic or the definitions of a first-order structure or of a deductive system.

We utilize the concept of metalogic to resolve this circularity:

• Using the metatheory where we assume the availability of first-order logic, we build in the object logic certain special sets within set theory that can be used as models of ZFC — see theorem 1104. In particular, the theorem implies that the existence of a regular strong limit cardinal in the metatheory can prove that ZFC is consistent.

• We declare such a set as “the” model of ZFC we are interested in. At this point, this set that we have built in the object logic becomes the ambient universe in the metatheory.

• Now we can, within the metatheory, define first-order logic.

An important point is that we should restrict ourselves to standard transitive models in order to avoid very counterintuitive results.

Another important point is that it is possible that the set theory which we use within the metatheory in order to provide models of ZFC is itself inconsistent. In that case, due to (AX EFQ), every theorem can be derived and a proof of consistency of ZFC is insubstantial.

Note that we do need additional assumptions for the axiom of universes separately because we cannot build a model using the cumulative hierarchy.

Remark 922. Instead of discussing first-order theories like the theory of groups, we can instead reformulate the definition within set theory and add a predicate formula IsGroup[\(\xi\)] with which is valid only for groups. That is, IsGroup[\(\xi\)] is a tautology if and only if the variable \(\xi\) satisfies the set-based definition of a group given in definition 455.
This is a natural approach, and we usually assume it implicitly. Furthermore, it makes no sense to speak about concepts like the lattice of subgroups or the cardinality of a group otherwise.

This is also a natural framework for defining topological spaces and categories via quivers. Some theories like the partially ordered sets are first-order theories, however well-ordered sets is an extension that requires an ambient set theory.

Thus, roughly, set theory allows us to use higher-order relations and types in first-order logic.
13.1. Naïve set theory

Naïve set theory is traditionally defined informally by only specifying that a set is an unordered collection of objects without repetition. It turns out that this can easily be formalized as a first-order theory, albeit an inconsistent one. Still, this theory is useful for introducing important concepts that can ease the introduction of more elaborate theories like ZFC. The definitions we introduce and the proofs we provide will turn out to be valid in ZFC also. In other words, we will transparently utilize unrestricted comprehension for constructing sets and later in proposition 985 we will prove that they exist not only in naïve set theory, but also in ZFC.

Remark 923. What we lose in this formalization are objects which are not sets, usually called atoms or urelements (because of the German prefix “ur”, meaning primordial). It is not necessary for us to add such elements since we can encode everything via sets. Theories without atoms, like our versions of naïve set theory and ZFC+U, are called pure set theories.

Definition 924. The language of naïve set theory is a first-order language $\mathcal{L}$ with only a single infix binary relation $\in$ called set membership. If $\xi \in \eta$, we say that $\xi$ is a member or element of $\eta$ and that $\eta$ contains $\xi$.

For the sake of simplicity, we will not introduce into the language any other functional or predicate symbols, but will use predicate formulas when needed mostly for formulating axioms. See the IsUniverse[$\tau$] predicate for an extreme example.

Naïve set theory is a first-order theory axiomatized by the following:

(a) The axiom of extensionality, which states that two sets are equal if and only if they have the same members. Symbolically,

$$\left( \forall \xi \left( \xi \in \tau \iff \xi \in \sigma \right) \right) \rightarrow \left( \tau \equiv \sigma \right).$$

As a consequence, a set is only distinguished by what it contains and thus the ordering and repetition of members of a set play no role. This axiom is also important in ZFC — see definition 984 (a).

It is very common when dealing with sets, as in (321), to use relativization of quantifiers with $\in$.

As explained in remark 781 (g), we avoid excessive universal quantification. We actually add as an axiom of the theory the universal closure of (321):

$$\forall \tau \cdot \forall \sigma \cdot \left( \left( \forall \xi \left( \xi \in \tau \iff \xi \in \sigma \right) \right) \rightarrow \left( \tau \equiv \sigma \right) \right).$$

The converse of (321) obvious.

(b) The axiom schema of unrestricted comprehension states that any formula defines a set. For each formula $\varphi$ not containing $\tau$ as a free variable, the following is an axiom:

$$\exists \tau \cdot \forall \xi \cdot (\xi \in \tau \iff \varphi).$$
It is important to highlight that \( \varphi \) may have any number of free variables as long as \( \tau \) is not among them. Of course, this axiom is only interesting if \( \xi \in \text{Free}(\varphi) \). If \( \eta_1, \ldots, \eta_n \) are all the other free variables of \( \varphi \), then the universal closure of the corresponding axiom is

\[
\forall \eta_1 \cdots \forall \eta_n . \exists \tau . \forall \xi . (\xi \in \tau \leftrightarrow \varphi).
\]  

(324)

In other words, the set \( \tau \) is not unique in general, but depends on the free variables \( \eta_1, \ldots, \eta_n \). For this reason, they are called parameters of the axiom.

Compare this axiom schema to restricted comprehension. In the context of naïve set theory they are equivalent because each is a special case of the other one.

Because our goal is for all our constructions to be valid in ZFC, we only use unrestricted comprehension where the existence of the set is justified by other axioms of ZFC.

**Remark 925.** The symbol \( \in \) is derived from \( \varepsilon \). Some older books like [Kel55] even use \( \varepsilon \) for set membership. Theorem 998 (Epsilon induction) is named after set membership.

**Definition 926.** Assume that we have a fixed standard transitive model \( \mathcal{V} = (V, I) \) of naïve set theory or ZFC, with or without the axiom of universes. We will assume ZFC+U by default.

We say that any member of \( V \) is a set. If \( A \) is a set and \( x \in A \), we say that \( x \) is a member or element of \( A \) or, in accordance with definition 403, a point in \( A \).

We usually refer to \( V \) as our universe or universal set. When working with Grothendieck universes, we may wish to further distinguish between the universal set and some Grothendieck universe. Fortunately, we very rarely refer to \( V \) itself.

**Remark 927.** We will say that a model \( \mathcal{V} = (V, I) \) of set theory is standard if the interpretation \( I(\in) \) of the membership predicate symbol is precisely the membership relation in the metatheory. We will only consider standard models of set theory. This is immensely important for the following reasons:

- Set-builder notation relies on constructing sets in the metatheory and then using them in the object theory. If the model is not standard, then it does not hold that \( \langle \xi \in \eta \rangle [\xi \mapsto x, \eta \mapsto y] = T \) if and only if \( x \in y \).
- It is possible that cardinality is incompatible between the object theory and metatheory — see example 911 (Skolem’s paradox). We want to avoid countable sets in the metatheory to be uncountable in the object theory, for example.

Therefore, it is reasonable to assume that all our models of set theory are standard. We also want their domains to be transitive sets – see remark 946.

**Definition 928.** As mentioned in remark 921, set theory somewhat blurs the line between logic and metalogic. In particular some definable subsets of the universe \( U \) of the fixed model \( \mathcal{U} \) are themselves sets within the object logic.

Fix a formula \( \varphi \) whose free variables are \( \xi \) and \( \eta_1, \ldots, \eta_n \). In the simplest case, \( n = 0 \) and \( \xi \) is the only free variable of \( \varphi \).
Also fix an $n$-tuple $u_1, \ldots, u_n$ of members of $U$, which we will call **parameters**. Denote by $A$ the subset of $U$ consisting of members $x$ of $U$ such that $\varphi[x, u_1, \ldots, u_n] = T$.

We introduce a special convenience notation for $A$ called **set-builder notation**:

$$A := \{x \mid \varphi[x, u_1, \ldots, u_n]\}.$$

Since set-builder notation is metalogical, we do not impose strict syntax rules and use prose where it is straightforward to translate it into a logical formula.

For example, the intersection of the sets $B$ and $C$ is given by the formula $\xi \in \eta \land \xi \in \zeta$, where $B$ is a value for the parameter $\eta$ and $C$ is a parameter for $\zeta$. The intersection can thus be written as

$$A := \{x \mid x \in B \land x \in C\}.$$

Note that at this point $A$ is a set within the metatheory and its members are sets within the object logic, however $A$ may not be a set within the object logic and its members may not be sets within the metatheory.

Nonetheless, within naïve set theory, as a consequence of the **axiom schema of unrestricted comprehension**, $A$ is also a set within the object logic. More precisely, given our choice of parameters $u_1, \ldots, u_n$, the axiom schema instance (324) guarantees the existence of a member $a$ of $U$, such that

$$(\xi \in \tau)[\tau \mapsto a, \xi \mapsto x] = T \text{ if and only if } x \in A,$$

where we have denoted set membership within the object logic by $\in$ and within the metatheory by $\vDash$. We will not further use this symbol and the two membership relations will be used interchangeably.

This is where the line between logic and metalogic blurs — we can speak about roughly the same sets within the object logic and the metatheory.

Examples such as theorem 932 (Russell’s paradox) show that unrestricted comprehension can easily lead to an inconsistent object logic. In more elaborate set theories like **ZFC**, we only allow restricted comprehension via the **axiom schema of specification**. Instead of defining $A$ as a set of all members of $U$ satisfying a certain property, restricted comprehension allows us to define $A$ as a subset not of the universe $U$, but of some well-behaved subset $B$ of $U$. The corresponding notation is

$$\{x \in B \mid \varphi[x, u_1, \ldots, u_n]\}.$$

Of course, we may still use unrestricted comprehension of the result is guaranteed to be a set within the object logic.

Within **ZFC**, subsets of $U$ which are not sets in the object logic are called **proper classes**. Sets and proper classes are collectively called **classes**. We avoid referencing proper classes because that can easily lead us to an inconsistent theory. See definition 1112 for a clever workaround.

A bigger problem that may happen is described in remark 946.

Other liberties regarding set-builder notation include the following:
• We often place arbitrary terms on the left side rather than only sets. This is simply a convenient metalogical notation; the symbols that are used in these terms are often not part of the object language. For example, we write the odd integers as

\[ \{2n + 1 \mid n \in \mathbb{Z} \}. \]

• Instead of using the delimiter |, we sometimes also use :, especially when dealing with absolute values and divisibility:

\[ \{ |n| : n \mid 125 \} \]

can be more readable than

\[ \{ |n| \mid n \mid 125 \} \]

• If a set has only a small finite amount of members, we usually prefer to enumerate them as

\[ \{1, 3, 9, 27\}. \]

Because of the axiom of extensionality, the order and repetition of the objects inside the curly braces are irrelevant. Nevertheless, using any unconventional order does not benefit us in any way.

• We can also place an ellipsis if a certain pattern is obvious:

\[ \{1, 3, 9, 27, \ldots\}. \]

This works specifically for defining countable sets.

Note that we have used certain numbers, but this was only for illustrative purposes because even the natural numbers are not yet defined in terms of sets.

**Remark 929.** Within the metatheory, we often use the notation \( x_1, \ldots, x_n \in A \) to mean that \( x_k \in A \) for \( k \in 1, \ldots, n \).

**Remark 930.** Sets with a single elements are usually called **singletons**. It is sometimes convenient, especially with connection to geometry or multi-valued functions (e.g. when dealing with limits of nets or subdifferentials), to not distinguish between singleton sets and their corresponding element.

**Definition 931.** A very important set is the **empty set**

\[ \emptyset := \{ x \mid \bot \}, \]

which contains no elements. We will also find useful the **predicate formula**

\[ \text{IsEmpty}[\tau] := \forall \eta. \neg \eta \in \tau. \]

We will often refer to **nonempty sets**, which are exactly what they sound — sets that are not the empty set.
**Theorem 932** (Russell’s paradox). **Naïve set theory is inconsistent.** More precisely, the instance of the schema of unrestricted comprehension with

\[ \varphi = (\xi \not\in \xi) \]  

allows us to derive \( \bot \) in classical logic.

Thus, the set

\[ R := \{ x \mid x \not\in x \} \]  

of all sets that do not contain themselves is not well-defined. Indeed, from \( R \not\in R \) it follows that \( R \in R \) and from \( R \in R \) it follows that \( R \not\in R \).

**Proof.** After substituting \( (325) \) in \( (323) \), we obtain the following axiom of naïve set theory:

\[ \psi := \exists \tau . \forall \xi . (\xi \in \tau \leftrightarrow \neg(\xi \in \xi)). \]  

(327)

We will show that the negation \( \neg \psi \) of \( \psi \) is also derivable in this theory. An explicit form of the negation can be obtained by utilizing the equivalences **proposition 842** and **proposition 815 (i)**:

\[ \neg \psi = \forall \tau . \exists \xi . (\xi \in \tau \leftrightarrow \xi \in \xi). \]

This holds when \( \xi \) and \( \tau \) take on the same value, hence it is satisfiable in naïve set theory. By **theorem 904**, \( \neg \psi \) is derivable in the theory.

Thus, \( \psi \) and \( \neg \psi \) are both derivable in the same theory, and we can use \( (\neg \neg) \) to also derive \( \bot \), which shows that naïve set theory is inconsistent. \( \square \)

**Definition 933.** We say that \( A \) is a **subset** of \( B \) and write \( A \subseteq B \) if every member of \( A \) is a member of \( B \). If \( A \) is a subset of \( B \), we say that \( B \) is a **superset** of \( A \).

If \( A \subseteq B \) and \( A \neq B \), we say that \( A \) is a **proper subset** of \( B \) and write \( A \subset B \).

The relation \( \subseteq \) is called the inclusion relation, and it gives a partial ordering between sets. See **proposition 938**. If an entire family of sets are not pairwise comparable, we say that they are **disjoint**.

The following **predicate formula**

\[ \text{IsSubset}[\rho, \tau] := \forall \xi . (\xi \in \rho \rightarrow \xi \in \tau), \]

which is valid when \( \rho \) is a subset of \( \tau \), will occasionally be useful for us. We have chosen the letter \( \rho \) because it is the result that would be on the right-hand side if we had a corresponding infix relation in the language.

**Remark 934.** Some authors, such as [Kel55], use the notation \( A \subset B \) to mean “all elements of \( A \) belong to \( B \)”, even in the case when \( A = B \). To avoid confusion, we use the notations \( A \subseteq B \) and \( A \subset B \).

**Remark 935.** In a pure set theory, everything is encoded as a set. However, it is often the case that we are not interested in how a set’s elements are encoded as sets and only in how they behave, e.g. when working with natural numbers, we are interested in the elements of \( \mathbb{N} \) and not in the way every element of \( \mathbb{N} \) is encoded as a set.
In order to reduce repetitiveness, sets whose elements we consider to be other sets, are often called families of sets. In particular, if all (different) sets are disjoint, we say that the family is a disjoint family. It is usually assumed that the sets are nonempty.

We often consider indexed families, i.e. sets which depend on a parameter, which further highlight our intention to distinguish between a point in a set, the set itself and some family of sets to which the latter belongs.

**Definition 936.** We define the following operations:

(a) Dually to intersections, the **union** of an arbitrary set $\mathcal{A}$ is defined as
\[
\bigcup \mathcal{A} := \{x \mid \exists A \in \mathcal{A}. x \in A\}.
\]

We define the predicate formula
\[
\text{IsUnion}[\rho, \tau] := \forall \xi. (\xi \in \rho \leftrightarrow \exists \eta \in \tau. \xi \in \eta).
\]

In particular, $\bigcup \emptyset = \emptyset$.

For two sets $A$ and $B$, we define the **binary union** as
\[
A \cup B := \bigcup \{A, B\} = \{x \mid x \in A \text{ or } x \in B\}.
\]

(b) The **intersection** of a nonempty set $\mathcal{A}$ is
\[
\bigcap \mathcal{A} := \{x \mid \forall A \in \mathcal{A}. x \in A\}.
\]

We also introduce the predicate formula
\[
\text{IsIntersection}[\rho, \tau] := \forall \xi. (\xi \in \rho \leftrightarrow \forall \eta \in \tau. \xi \in \eta).
\]

We leave $\bigcap \emptyset$ undefined because it should be a top element in the Boolean algebra of all sets, but the latter object is an ambiguous object and does not even exist in ZFC — see proposition 985 (c). It does nonetheless satisfy $\text{IsIntersection}[\rho, \tau]$.

For two sets $A$ and $B$, we define the **binary intersection** as
\[
A \cap B := \bigcap \{A, B\} = \{x \mid x \in A \text{ and } x \in B\}.
\]

(c) The **difference** of the sets $A$ and $B$ is
\[
A \setminus B := \{x \in A \mid x \notin B\}.
\]

We define the predicate formula
\[
\text{IsDifference}[\rho, \tau, \sigma] := \forall \xi. (\xi \in \rho \leftrightarrow \xi \in \tau \land \neg(\xi \in \sigma)).
\]
(d) The **power set** \( \text{pow}(A) \) of \( A \) is the family of all subsets of \( A \). Symbolically,

\[
\text{pow}(A) := \{ B \mid B \subseteq A \}.
\]

The operation \( \text{pow} \) is not technically a function since its domain is supposed to be the set of all sets, whose existence contradicts theorem 932 (Russell’s paradox). Nevertheless, this notation makes sense and is justified by remark 1054 (Unbounded transfinite recursion) and example 1136.

We define the predicate formula

\[
\text{IsPowerSet}\{\rho, \tau\} := \forall \xi. \left( \xi \in \rho \leftrightarrow \text{IsSubset}\{\xi, \tau\} \right).
\]

See lemma 1085 for a characterization of the power set.

**Proposition 937.** *Set difference* has the following basic properties:

(a) If \( A \) and \( B \) are subsets of \( C \), then \( A \setminus B = A \cap (C \setminus B) \).

(b) If \( A \subseteq B \), then \( B \setminus (B \setminus A) = A \)

**Proof.**

**Proof of 937 (a).** Since \( a \in A \) implies \( a \in C \), we have

\[
A \setminus B = \{ x \in A \mid x \notin B \} =
\]

\[
= \{ x \in A \mid x \in C \text{ and } x \notin B \} =
\]

\[
= A \cap (C \setminus B).
\]

**Proof of 937 (b).** By double negation elimination,

\[
B \setminus (B \setminus A) = \{ x \in B \mid x \notin \{ x \in B \mid x \notin A \} \} =
\]

\[
= \{ x \in B \mid x \in A \} =
\]

\[
= A.
\]

\[\square\]

**Proposition 938.** Let \( X \) be an arbitrary set. Then the power set \( \text{pow}(X) \) endowed with the inclusion partial order \( \subseteq \) is a complete Boolean algebra. Explicitly:

(a) The *join* of an arbitrary family \( A \) of subsets of \( X \) is simply the union \( \bigcup A \).

(b) The *top element* is the set \( X \) itself.

(c) The *meet* of an arbitrary family \( A \) of sets is simply the intersection \( \bigcap A \). Unlike for a general family of sets, we have no problem defining the intersection of an empty set to be the top element \( X \).
Figure 27: The Hasse diagram of pow({A, B}) with respect to set inclusion

(d) The bottom element is the empty set.

(e) The complement $A^c$ of the subset $A$ is the difference $X \setminus A$.

Proof.

Proof of 938 (a). The union of $A$ exists by proposition 985 (g), and it is itself a subset of $A$. Every set in $A$ is contained in $\bigcup A$, hence it is indeed a join.

Proof of 938 (b). Clearly $X$ contains every subset of $X$.

Proof of 938 (a). The intersection of $A$ exists by proposition 985 (g), and it is itself a subset of $A$. Every set in $A$ contains $\bigcap A$, hence it is indeed a meet.

Proof of 938 (d). The empty set is contained in every set, in particular in every subset of $A$.

Proof of 938 (e). The operation $A^c$ is well-defined for each subset $A$ of $X$ due to proposition 985 (i).

By definition

$$A \lor A^c = A \cup (X \setminus A) = X$$

and

$$A \lor A^c = A \cup (X \setminus A) = X,$$

hence $A^c$ is indeed the complement of $A$.

Therefore, pow($X$) is a Boolean algebra. \qed

Remark 939. Induction is an important proof technique that is discussed in detail in the proof of proposition 5. There are more general forms of induction than (PA3) like theorem 997 (Well-founded induction) and theorem 997 (Well-founded induction). They do, however, require concepts which in turn depend on the existence of natural numbers within set theory. As a consequence, we cannot prove (PA3) via theorem 997 (Well-founded induction).

We will introduce the concept of inductive sets in definition 941 and prove in theorem 943 (Induction via inductive sets) that a special inductive set $\omega$, which will be the domain of our model of $\mathbb{N}$, allows performing inductive proofs. The technique that allows us to perform inductive proofs on $\omega$ can be seen in the proof of proposition 945. Theorem 943 (Induction via inductive sets) will allow us to define natural numbers without relying on metalogical induction along the way. See the proof of theorem 943 (Induction via inductive sets) for
a description of natural number induction within set theory and remark 4 for a further discussion of the use of natural numbers in the metatheory and in the object logic.

We also introduce recursion in parallel as a technique for constructing objects. See theorem 979 (Recursion theorem).

**Definition 940.** The successor $\text{succ}(A)$ of a set $A$ is the set

$$\text{succ}(A) := A \cup \{A\}.$$  

It is also called the ordinal successor operation since it is an important concept in the theory of ordinals. See remark 1022 for an example of how it naturally arises. It should be distinguished from successor cardinals defined in definition 1050.

The following predicate formula

$$\text{IsSucc}[\rho, \tau] := \forall \xi. \left( \xi \in \rho \iff (\xi \in \tau \lor \xi = \tau) \right),$$

which states that $\rho$ is the successor of $\tau$, will be useful for us when working with inductive sets.

**Definition 941.** A set is called inductive if contains the empty set and is closed under the successor operator.

We introduce the following predicate formula

$$\text{IsInductive}[\tau] := \exists \xi \in \tau. \text{IsEmpty}[\xi] \land \left( \forall \xi \in \tau. \exists \eta \in \tau. \text{IsSucc}[\eta, \xi] \right).$$

**Proposition 942.** There is a smallest (with respect to set inclusion) inductive set, which we denote by $\omega$.

**Proof.** We cannot directly define $\omega$ as the intersection of all inductive sets since we want to avoid unrestricted comprehension. Fortunately, the existence of at least one inductive set $A$ is justified by the axiom of infinity in ZFC or by taking the entire universe in naïve set theory.

Hence, we use restricted comprehension:

$$\omega := \{x \in A \mid x \text{ belongs to every inductive set}\}.$$  

To see that $\omega$ is itself inductive, note that $\emptyset \in \omega$ and that if $x \in \omega$, then it also belongs to all inductive sets and hence $\text{succ}(x)$ also belongs to all inductive sets, proving $\text{succ}(x) \in \omega$. □

**Theorem 943** (Induction via inductive sets). We can perform induction on the smallest inductive set $\omega$. That is, we can prove that some property holds for every element of $\omega$ by proving the following:

- The property holds for $\emptyset$
- We can prove that it holds for $\text{succ}(n)$ by assuming that it holds for some set $n \in \omega$.
This is an analog of (PA3) and is actually used in theorem 981 to prove that \( \omega \) is a model of \( PA \). Instead of an entire theorem schema, however, for this theorem it is sufficient to use one single formula. The more general induction principles that use theorem schemas cannot be proved without natural numbers, which are a model of \( PA \) by virtue of this theorem.

More formally, the following is a theorem of both naïve set theory and \( ZF \):

\[
\exists \sigma \forall \tau. \left( \left( \exists \xi \in \tau. \text{IsEmpty} [\xi] \right) \land \forall \xi. \left( \xi \in \tau \rightarrow \exists \eta \in \tau. \text{IsSucc} [\eta, \xi] \right) \right) \rightarrow \text{IsSubset} [\sigma, \tau] \tag{328}
\]

Proof. The antecedent of (the inner formula in) (328) is a restatement of the predicate formula \( \text{IsInductive} [\tau] \). The situation resembles the axiom of infinity, but, instead of existence of an inductive set \( \tau \), it states the existence of a set \( \sigma \) such that if \( \tau \) is an inductive set, then \( \sigma \) is a subset of \( \tau \) (if we restrict \( \xi \) to range only over members of \( \sigma \), then we would obtain equality of \( \tau \) and \( \sigma \) instead). In other words, we have reduced the verification of (328) to showing that there exists a minimal inductive set in both naïve set theory and \( ZF \).

We have already proved in proposition 942 that our fixed model \( \mathcal{V} = (V, I) \) of set theory has a minimal inductive set \( \omega \). Thus, for any variable assignment \( v : \text{Var} \rightarrow V \), the modified assignment \( v_{\exists \omega} \omega \) satisfies (328) with the outer existential quantifier removed. Hence, by definition 828 (d), it follows that the entire formula (328) is satisfied by the assignment \( v \).

Both the assignment \( v \) and the model \( \mathcal{V} \) were arbitrary, therefore we can conclude that (328) is a theorem of both naïve set theory and \( ZF \).

Definition 944. A set \( A \) is transitive if from \( B \in A \) it follows that \( B \subseteq A \).

See remark 1002 for a discussion of the motivation and terminology of transitive sets and remark 946 for their importance.

We introduce the following predicate formula:

\[
\text{IsSetTransitive} [\tau] := \forall \xi \in \tau. \forall \eta \in \xi. \eta \in \tau
\]

Proposition 945. The set \( \omega \) is transitive and every member of \( \omega \) is transitive.

This proof demonstrates usage of theorem 943 (Induction via inductive sets).

Proof.

Proof that all members of \( \omega \) are transitive. In order to demonstrate how theorem 943 (Induction via inductive sets) works in practice, we will use inductive sets directly. Let \( T \subseteq \omega \) be the subset of all transitive members of \( \omega \). We will show that \( T \) is inductive. Since \( T \subseteq \omega \) and \( \omega \) is the smallest inductive set, we can conclude that \( T = \omega \).

Clearly \( \emptyset \in T \) because every member of \( \emptyset \) vacuously is a subset of \( \emptyset \).

Now suppose that \( n \in T \) and let \( m \in \text{succ}(n) = n \cup \{n\} \). If \( m = n \), then \( m \in \text{succ}(n) \) by definition of the successor operation. If \( m \in n \), then \( m \subseteq n \) by the inductive hypothesis and hence also \( m \subseteq n \cup \{n\} = \text{succ}(n) \). Thus, \( \text{succ}(n) \) is also transitive.

We have shown that \( T \) is inductive. Therefore, \( \omega = T \) and every member of \( \omega \) is transitive. From now on, we will not be as explicit about the use of induction on \( \omega \).
**Proof that \( \omega \) is transitive.** We will show that for all members \( n \) of \( \omega \) we have \( n \subseteq \omega \).

The case \( n = \emptyset \) is again trivial.

Now suppose that \( n \subseteq \omega \) and let \( m \in \text{succ}(n) \). If \( m = n \), clearly \( m \subseteq \omega \). If \( m \in n \), then \( m \subseteq n \) and, since \( n \subseteq \omega \), we have \( m \subseteq \omega \) by transitivity of \( \subseteq \).

Therefore, \( \omega \) is transitive. \( \square \)

**Remark 946.** As discussed in definition 928, within the axiom schema of unrestricted comprehension it may happen that \( U \subseteq V \) is not a set within the object logic.

But there is a bigger problem that may happen even for standard models. If \( A \in V \) and \( x \in A \) (in the metatheory), it is possible that \( x \) is not in \( V \). Therefore, if we have shown that \( A \) is a set within the object logic, it is possible that its members within the metatheory are not members in the object logic. On other words, it is possible for set membership itself to be incompatible between the metatheory and object logic.

If \( V \) is a transitive set, however, we would not have such a problem. That is, if we construct a set \( A \) in the metatheory and show that it belongs to some set \( B \) in the object logic, then \( A \) itself would also be a set in the object logic.

For this reason, it is very important to consider only transitive models of set theory.

**Lemma 947.** No element of \( \omega \) is a member of itself.

**Proof.** We will again use theorem 943 (Induction via inductive sets). By definition, \( \omega \notin \omega \).

Now suppose that \( n \notin n \) for some \( n \in \omega \).

Aiming at a contradiction, suppose that \( \text{succ}(n) \in \text{succ}(n) \). The assumption that \( n = \text{succ}(n) = n \cup \{n\} \) implies that \( n = \{n\} \). The assumption that \( \text{succ}(n) \in n \) implies that \( n \in n \) since \( \text{succ}(n) \) is transitive by proposition 945. In both cases we have \( n \in n \), which contradicts our inductive hypothesis. This contradiction shows that \( \text{succ}(n) \notin \text{succ}(n) \).

Theorem 943 (Induction via inductive sets) allows us to conclude that no member of \( \omega \) contains itself. \( \square \)

**Theorem 948.** The smallest inductive set \( \omega \) satisfies the axioms (PA1)-(PA3) from Peano arithmetic with the following interpretation:

(a) Zero is interpreted as \( \emptyset \).

(b) The successor operation \( s \) is interpreted as \( \text{succ} \).

We will generalize this theorem to theorem 981 after we are able to define the arithmetic operations in \( \omega \).

**Proof.**

**Proof of PA1.** Let \( n, m \in \omega \) and suppose that \( \text{succ}(n) = \text{succ}(m) \). If \( n = m \), there is nothing to prove.

Suppose that \( n \neq m \). Thus, since

\[ n \cup \{n\} = \text{succ}(n) = \text{succ}(m) = m \cup \{m\}, \]

we have both \( n \in m \) and \( m \in n \).

Proposition 945 implies that \( n \) is transitive and hence \( n \notin n \), which contradicts lemma 947.

The obtained contradiction shows that \( n = m \).
Proof of **PA2.** Suppose that $\emptyset$ has a predecessor $n \in \omega$. Then

$$\emptyset = \text{succ}(n) = n \cup \{n\},$$

which implies that $n \in \emptyset$. But this contradicts the definition of $\emptyset$.

Therefore, $\emptyset$ has no predecessor.

Proof of **PA3.** It follows from theorem 943 (Induction via inductive sets) that (328) is a theorem of ZF. Let $\mathcal{V} = (V, I)$ be our ambient standard transitive model of ZFC.

Fix any variable assignment $v : \text{Var} \rightarrow V$. As in the proof of theorem 943 (Induction via inductive sets), we consider the modified assignment $v_{\sigma \mapsto \omega}$ that “eliminates” the outer existential quantifier in (328).

To show that (328) really corresponds to (PA3) (and hence that $\omega$ satisfies (PA3)), fix some formula $\phi$ of Peano arithmetic (not ZFC!) and suppose that $\xi, \zeta_1, \ldots, \zeta_n$ are all of its free variables. Fix also some parameter values $u_1, \ldots, u_n \in \omega$ and, as in definition 928, define the set

$$A := \{x \in \omega \mid \phi[x, u_1, \ldots, x_n] \}.$$

Since (328) is satisfied by $v$, the inner formula in (328) (without the quantifiers over $\sigma$ and $\tau$) is satisfied by $v_{\sigma \mapsto \omega, \tau \mapsto A}$.

Since our choice of parameters $u_1, \ldots, u_n$ was arbitrary, we can conclude that the universal closure (PA3') of (PA3) is satisfied by $\omega$ for every formula $\phi$ of PA.

Remark 949. Due to theorem 948, we will henceforth identify the smallest inductive set $\omega$ with the set $\mathbb{N}$ of natural numbers.

We are not yet able to add or multiply natural numbers, nor rely on their well-foundedness, however for all other purposes we are able to utilize them.

Since the ordering in definition 10 is defined via addition, we must define some other ordering. Luckily, as we shall see in section 13.6 (Ordinals), $n < m$ corresponds to $n \in m$. In particular, the members of $m$ are ordered.

As a consequence, every natural number $n$ equals the set of all smaller natural numbers by the axiom of extensionality. It is conventional to write $\{0, 1, \ldots, n\}$ rather than $\{m \mid m \in n\}$. The former notation will be fully justified in section 13.6 (Ordinals).

This is useful, for example, in definition 951.
13.2. Relations

Definition 950. We are now in a vicious cycle where we need binary Cartesian products in order to define arbitrary Cartesian products. We will do this as quickly as possible, without introducing relations and functions. The latter two will be discussed in detail in section 13.2 (Relations) and section 13.3 (Functions), respectively.

(a) The Kuratowski pair or simply ordered pair \((x, y)\) of the sets \(x\) and \(y\) is defined as

\[
(x, y) := \{\{x\}, \{x, y\}\}.
\]

This is a simple and widespread definition that encodes the order of \(x\) and \(y\), unlike the set \(\{x, y\}\) for example.

We will later use the notation \((x, y)\), but until remark 952, we want to distinguish between Kuratowski pairs and 2-tuples.

We will use the following predicate formula in IsFun\(\rho, \tau, \sigma\):

\[
\text{IsPair}\[\rho, \tau, \sigma\] := \forall \xi. \left(\xi \in \rho \leftrightarrow \left(\forall \eta \in \xi. \eta \doteq \tau \right) \lor \left(\forall \eta \in \xi. \left(\eta \doteq \tau \lor \eta \doteq \sigma\right)\right)\right)
\]

(b) A set \(i\) of Kuratowski pairs is called an indexed family if whenever \((k, A) \in i\) and \((k, B) \in i\), we have \(A = B\). It is conventional to denote this unique set corresponding to \(k\) as \(A_k\) without an explicit reference to \(i\).

The index set of the family is

\[
\mathcal{K} := \{k \mid \exists A. (k, A) \in i\}.
\]

The family itself is then denoted as

\[
\{A_k\}_{k \in \mathcal{K}}.
\]

If \(A_k \in \mathcal{A}\) for every \(k \in \mathcal{K}\), we sometimes write

\[
\{A_k\}_{k \in \mathcal{K}} \subseteq \mathcal{A},
\]

although the latter is an embedding rather than set inclusion.

(c) A tuple from the indexed family \(\{A_k\}_{k \in \mathcal{K}}\) is another indexed family \(\{x_k\}_{k \in \mathcal{K}}\) with the same index set satisfying the condition that for every \(k \in \mathcal{K}\), the value \(x_k\) belongs to \(A_k\).

We will see later that this is precisely a choice function.
The Cartesian product of an indexed family \( \{ A_k \}_{k \in \mathcal{K}} \) is the set of all tuples from this family. We denote the Cartesian product by \( \prod_{k \in \mathcal{K}} A_k \).

With the availability of functions in section 13.3 (Functions), we can define the Cartesian product as the set
\[
\prod_{k \in \mathcal{K}} A_k = \{ f : A \to \bigcup_{k \in \mathcal{K}} A_k | \forall k \in \mathcal{K} . k \in A_k \}.
\]

**Definition 951.** Families indexed by \( \omega \) are called infinite sequences or simply sequences. We will use several notations, depending on the context.

- \( \{ A_k \}_{k \in \mathbb{N}} \), which is the conventional notation for indexed families.
- \( \{ A_k \}_{k=0}^\infty \), which easily extends to cases such as \( \{ A_k \}_{k=m}^n \) when the index set is \( \{ m, ..., n \} \).
- \( (A_0, A_1, ...) \), which is used when explicitly enumerating members of the sequence.

It is conventional to write a family \( \{ A_k \}_{k \in \mathbb{N}} \) indexed by a natural number \( n \) using the notation \( (A_1, \cdots, A_n) \), with or without the outer parentheses. Families indexed by natural numbers are called finite sequences.

For \( n = 2 \), finite sequences are called pairs, for \( n = 3 \) — triples and for \( n = 4 \) — quadruples.

We say that \( A_{k_1}, A_{k_2}, ... \) is a subsequence of \( A_1, A_2, ... \) if \( k_1 < k_2 < ... \), i.e. if the sequence of indices is monotone in the sense of (497).

**Remark 952.** Note that the tuple \( (A, B) \)
\[
(A, B) = \left\{ \{ \{0\}, \{0, A\} \}, \{ \{1\}, \{1, B\} \} \right\}
\]

is formally different from the Kuratowski pair
\[
\langle A, B \rangle = \left\{ \{ A \}, \{ A, B \} \right\}.
\]

This is one reason we hurried to define general Cartesian products — we wanted to avoid working with tuples defined in terms of Kuratowski pairs. We even introduced a special notation for them, just so we can avoid any confusion. Nevertheless, it is conventional to conflate Kuratowski pairs with \( \{0,1\} \)-indexed tuples.

We also conflate the tuples \( (A, (B, C)), ((A, B), C) \) and \( (A, B, C) \).

**Definition 953.** Let \( A_1, ..., A_n \) be a finite sequence of sets and let
\[
R \subseteq A_1 \times \cdots \times A_n
\]
be a subset of their Cartesian product.

The sequence \((R, A_1, \ldots, A_n)\) is called an \(n\)-ary relation. We say that the tuple \((x_1, \ldots, x_n) \in A_1 \times \cdots \times A_n\) is related with respect to \(R\) if \((x_1, \ldots, x_n) \in R\).

Relations are the semantical counterpart to first-order predicates and are equivalent to Boolean-valued functions, as discussed in remark 797.

We generalize only the following notions from binary relations:

(a) The set \(R\) of tuples is called the **graph** of the relation. In case the sequence \(A_1, \ldots, A_n\) is clear from the context, we can identify the relation \((R, A_1, \ldots, A_n)\) with its graph \(R\). We occasionally use the notation \(\text{gph}(R)\) for explicitly denoting the graph.

(b) The **signature** of the relation is the sequence \((A_1, \ldots, A_n)\). Obviously this definition only makes sense if we know what the signature is, either from the context or from the definition of the relation as the sequence \((R, A_1, \ldots, A_n)\) rather than only via its graph \(R\).

As a matter of fact, it is common to ignore the signature when defining relations — see e.g. [Kel55, p. 7] or [Aut20, def. 2.1]. If we do identify a relation only with its graph, however some notions like ranges and images coincide despite being different and other notions like function surjectivity make no sense.

Furthermore, two relations whose graphs are equal may have different signatures, which further highlights how important it is to distinguish between a relation and its graph.

(c) For some small values of \(n\), \(n\)-ary relations have established names:

- **Nullary** if \(n = 0\).
- **Unary** if \(n = 1\).
- **Binary** if \(n = 2\).
- **Ternary** if \(n = 3\).

This is not to be confused with **function arity** — functions are always binary relations.

(d) If all \(A_k\) for \(k = 1, \ldots, n\) are equal to the set \(A\), we say that \(R \subseteq A^n\) is a relation on \(A\).

**Definition 954.** An important special case of relations are **binary relations**. Given two sets \(A\) and \(B\), a binary relation between them is a triple \((R, A, B)\).

In addition to the terminology for **definition 953**, we also introduce the following terms:

(a) The relation is **empty** if its graph is the empty set, i.e. if no two elements are related.

It is important to highlight that the graphs of all empty relations are equal, but two empty relations are only equal if their signatures are.

(b) The **converse relation** of \(R\) is

\[
R^{-1} := \{(y, x) \mid (x, y) \in R\}.
\]
(c) If $A = B$, the **restriction** of $R$ to $X \subseteq A$ is the binary relation $(R|_X, X, X)$ is

$$R|_X := R \cap (X \times X) = \{(x, y) \in R \mid x \in X \text{ and } y \in X\}.$$  

We say that $R$ is an **extension** of $R|_X$.

(d) A special relation is the **diagonal relation** on a set $A$:

$$\Delta_A := \{(x, x) \mid x \in A\}.$$  

(e) Given two binary relations $R \subseteq A \times B$ and $T \subseteq B \times C$, we define their composition as

$$T \circ R := \{(x, z) \in A \times C \mid \exists y \in B . ((x, y) \in R \text{ and } (y, z) \in T)\}.$$  

Whenever $A = B$ and $R$ is simply a binary relation over $A$, the following are commonly used conditions that are often as axioms to some theory:

(f) $R$ is **reflexive** if $\Delta_A \subseteq R$, i.e. if every element of $A$ is related with itself.

The following formula is used as an axiom for nonstrict partial orders and entourages:

$$\forall \xi . (\xi R \xi). \quad (329)$$  

Note that we use remark 825 (a) notation in the latter case. Using either infix or prefix notation is actually a necessity since we do not actually have a concept of an ordered tuple in general (not set-based) first-order theories — see remark 825 (a).

(g) $R$ is **irreflexive** if $\Delta_A \cap R = \emptyset$, i.e. if no element of $A$ is related with itself.

The following formula is used as an axiom for strict partial orders:

$$\neg \exists \xi . (\xi R \xi). \quad (330)$$  

(h) $R$ is **symmetric** if $R = R^{-1}$.

The following formula is used as an axiom for equivalence relations, undirected graphs and entourages:

$$\xi R \eta \rightarrow \eta R \xi. \quad (331)$$  

(i) $R$ is **antisymmetric** if $R \cap R^{-1} = \Delta_X$.

The following formula is used as an axiom for partial orders:

$$(\xi R \eta \land \eta R \xi) \rightarrow \xi \approx \eta. \quad (332)$$  

(j) $R$ is **transitive** if $R = R \circ R$.

The following formula is used as an axiom for preorders:

$$(\xi R \eta \land \eta R \zeta) \rightarrow \xi R \zeta. \quad (333)$$  

476
(k) $R$ is **total** if any two member of $A$ are related.

The following formula is used as an axiom for nonstrict total orders:

$$\forall \xi . \forall \eta . (\xi R \eta \lor \eta R \xi).$$

Total relations are also called **connected** because of theorem 1186 (c).

(l) $R$ is **trichotomic** if every two elements of $A$ are either related or equal.

The following formula is used as an axiom for strict total orders:

$$\forall \xi . \forall \eta . (\xi R \eta \lor \eta R \xi \lor \eta = \xi).$$

**Example 955.** Binary relations are used in vastly different contexts:

- **Functions** are special binary relations.
- **Orders** are also special binary relations.
- **Directed graphs** are commonly defined as binary relations.
- **Entourages** are binary relations in **uniform spaces**.
- Relations are equivalent to Boolean-valued functions as shown in remark 797, and are often used for defining semantics of predicate symbols in first-order logic.

**Definition 956.** A binary relation on the set $A$ that is reflexive, symmetric and transitive is called an **equivalence relation**. In other words, an equivalence relation is a symmetric preorder.

We often denote equivalence relations via the symbol $\sim$.

(a) The **equivalence class** of $x \in A$, also called its **coset**, is the set

$$[x] := \{y \in A \mid x \sim y\}$$

of all elements of $A$ that are related to $x$.

(b) The **quotient set** of $A$ by $\sim$ is the set

$$A/\sim := \{[x] \mid x \in A\}.$$

If we have an easy way to choose a representative from each coset, then $A/\sim$ may be regarded as a subset of $A$. In general, this is not a subset relation by only an embedding.

(c) Using forward references to section 13.3 (Functions), we define the **canonical projection** as the function

$$\pi : A \to A/\sim$$

$$\pi(x) := [x].$$
If we have a fixed choice function \( c : A/\sim \to A \), we also define the canonical embedding
\[
\iota : A/\sim \to A \quad \iota([x]) := c(x).
\]

We sometimes have an obvious choice function, for example in proposition 642. In this case, the canonical projection may be regarded as a function from \( A \) to the subset \( c(A/\sim) \) of \( A \). Otherwise, the function \( \pi \) can be regarded as a multi-valued function from \( A \) to \( A \).

**Remark 957.** If \( x \sim y \) for some equivalence relation, we say that they are **congruent modulo \( \sim \)**. This concept specializes to congruence modulo normal subgroups defined in definition 515.

**Proposition 958.** The equality relation \( = \) is the intersection of all equivalence relations.

**Proof.** It is equivalent to the diagonal relation \( \Delta_X \). It is the smallest reflexive (resp. symmetric and transitive) relation on \( A \), i.e. the intersection of all reflexive (resp. symmetric and transitive) relations.

**Definition 959.** Let \( A \) be a set. A **cover** of \( A \) is a family \( \mathcal{A} \subseteq \text{pow}(A) \) of nonempty subsets of \( A \) such that \( A = \bigcup \mathcal{A} \). We sometimes use the term more loosely and say that an arbitrary family of sets \( \mathcal{A} \) is a cover of \( A \) if \( A \subseteq \bigcup \mathcal{A} \). The two definitions are identical if we intersect each set in \( \mathcal{A} \) with \( A \) and exclude the empty sets.

A **partition** of \( A \) is a pairwise disjoint cover. In other words, the cover \( \mathcal{A} \) is a partition if and only if each element of \( A \) belong to exactly one set in \( \mathcal{A} \).

**Proposition 960.** Fix a set \( A \). Let \( \sim \) be a binary relation of \( A \). The following are equivalent:

(a) \( \sim \) is an equivalence relation.

(b) There exists a partition \( \mathcal{A} \) of \( A \) such that \( x \sim y \) if and only if they belong to the same set in the partition \( \mathcal{A} \).

**Proof.**

**Proof that 960 (a) implies 960 (b).** Let \( \sim \) be an equivalence relation on \( A \). The quotient set \( A/\sim \) is a partition. Indeed:

- Every element \( x \in A \) belongs exactly one equivalence class \([x]\) by definition.

- The equivalence classes are disjoint. Indeed, assume the contrary. Then there exist \( x \) and \( y \) such that \([x] \cap [y] \neq \emptyset\) and yet \( x \sim y \).

  Let \( z \in [x] \cap [y] \). Then \( z \sim x \) and \( z \sim y \), thus from transitivity of \( \sim \) we have \( x \sim z \sim y \) and hence \( x \sim y \), which contradicts our assumption that \( x \sim y \).

  Hence, either \([x] = [x]\) or \([x] \cap [y] = \emptyset\). That is, different equivalence classes are disjoint.
Proof that 960 (b) implies 960 (a). Let \( \mathcal{A} \) be a partition of \( A \) such that \( x \sim y \) if and only if they both belong to the same set in \( \mathcal{A} \).

Given \( x \in A \), denote by \( A_x \) the set in \( \mathcal{A} \) which contains \( x \). The family \( \{ A_x \}_{x \in A} \) is well-defined since \( \mathcal{A} \) is a partition, which means that \( x \) belongs to exactly one set in \( \mathcal{A} \).

Proof of reflexivity. Clearly \( A_x = A_x \), hence \( x \sim x \).

Proof of symmetry. If \( x \sim y \), then \( A_x = A_y \), which implies \( A_y = A_x \) and thus \( x \sim y \).

Proof of transitivity. If \( x \sim y \) and \( y \in z \), then \( A_x = A_y = A_z \) and thus \( x \sim z \).

\[ \square \]

Definition 961. Let \( R \subseteq A^2 \) be a binary relation on the set \( A \). We define several closure operators:

(a) The reflexive closure of \( R \) is

\[ \text{cl}^R(R) := R \cup \Delta_X. \]

(b) The symmetric closure of \( R \) is

\[ \text{cl}^S(R) := R \cup R^{-1}. \]

(c) The transitive closure \( \text{cl}^T(R) \) of \( R \) is

\[ \text{cl}^T(R) := \bigcup \{ R^k \mid k = 1, 2, \ldots \}, \]

where \( R^k \) is iterated composition of \( R \).

Note that this is very different from the transitive closure of a set defined in definition 1016.

A transitive reduction of \( R \) is a minimal relation \( Q \subseteq R \) such that \( \text{cl}^T(Q) = \text{cl}^T(R) \). If there exists a smallest such relation, it is the unique transitive reduction, and we denote it by \( \text{red}^T(R) \).

Proposition 962. Binary relation closures have the following basic properties:

(a) The symmetric and transitive closures of a reflexive relation are symmetric.

(b) The reflexive and transitive closures of a symmetric relation are symmetric.

(c) The reflexive closure of a transitive relation is transitive. The symmetric closure of a transitive relation may not be transitive — see example 963.

(d) The reflexive and symmetric closures commute:

\[ \text{cl}^S \text{cl}^R(R) = \text{cl}^R \text{cl}^S(R). \]

\[ (336) \]

(e) The transitive and reflexive closures commute:

\[ \text{cl}^R \text{cl}^T(R) = \text{cl}^T \text{cl}^R(R). \]

\[ (337) \]
For the transitive and symmetric closures of $R$ we have
\[
\text{cl}^S \cdot \text{cl}^T (R) \subseteq \text{cl}^T \cdot \text{cl}^S (R).
\] (338)

The converse holds if $R$ is symmetric but not in general — see example 963.

Proof.

Proof of 962 (a). Trivial.
Proof of 962 (b). Trivial.
Proof of 962 (c). Trivial.
Proof of 962 (d). Trivial.
Proof of 962 (e). The reflexive closure only adds pairs of the form $(x, x)$. Thus, if $(x, y) \in \text{cl}^T (\text{cl}^R (R))$ for $x \neq y$, then $(x, y) \in \text{cl}^T (R) \subseteq \text{cl}^R (\text{cl}^T (R))$.
Conversely, if $(x, y) \in \text{cl}^R (\text{cl}^T (R))$ for $x \neq y$, then $(x, y) \in \text{cl}^T (R) \subseteq \text{cl}^R (\text{cl}^T (R))$.

Proof of 962 (f). If $(x, y) \in \text{cl}^S (\text{cl}^T (R))$, then we have the following possibilities:

- If $(x, y) \in R$, obviously $(x, y) \in \text{cl}^T (\text{cl}^S (R))$.
- If $(x, y) \in \text{cl}^T (R) \setminus R$, then there exists some natural number $k > 1$ such that $(x, y) \in R^k$.
  Since $R^k \subseteq [\text{cl}^S (R)]^k$, as can be shown by induction, we have $(x, y) \in [\text{cl}^S (R)]^k$. We thus conclude that $(x, y) \in \text{cl}^T (\text{cl}^S (R))$.
- Finally, if $(x, y) \notin \text{cl}^T (R)$, then $(y, x) \in \text{cl}^T (R)$. As in the previous step, we can show that $(y, x) \in \text{cl}^T (\text{cl}^S (R))$. The latter set is symmetric, hence $(x, y) \in \text{cl}^T (\text{cl}^S (R))$.

Since $(x, y)$ was arbitrary, we conclude that (338) holds.
Furthermore, if $R$ is symmetric, then
\[
\text{cl}^T \cdot \text{cl}^S (R) = \text{cl}^T (R) = 962 (a) = \text{cl}^S \cdot \text{cl}^T (R).
\]

Example 963. We will show that symmetric and transitive closures of relations do not commute. This is also a consequence of the difference between weak and strong connectedness of quivers.

Consider the set $A = \{a, b, c\}$ and the relation $R = \{(a, b), (c, b)\}$.

It should be noted that $R$ is transitive. Thus,
\[
\text{cl}^S (\text{cl}^T (R)) = \text{cl}^S (R) = R \cup \{(b, a), (b, c)\}.
\]

The latter set is not transitive because $(a, b)$ and $(b, a)$ both belong to $\text{cl}^S (R)$ and neither $(a, a)$ nor $(b, b)$ do not.
This shows that the converse of (338) does not hold in general.
**Proposition 964.** The reflexive, symmetric and transitive closure $\text{cl}^T \text{cl}^S \text{cl}^R(R)$ of any relation $R$ is an equivalence relation.

This holds for any permutation of the closures as long as $\text{cl}^T$ is applied after $\text{cl}^S$. This latter restriction is due to proposition 962 (f).

**Proof.** Let $R \subseteq A \times B$ be an arbitrary relation. By proposition 962 (a), $\text{cl}^S \text{cl}^R(R)$ is reflexive. It is also symmetric as the symmetric closure of $\text{cl}^R(R)$.

Then the transitive closure $\text{cl}^T \text{cl}^S \text{cl}^R(R)$ is also symmetric and reflexive by proposition 962 (a) and proposition 962 (b).

Therefore, $\text{cl}^T \text{cl}^S \text{cl}^R(R)$ is an equivalence relation. \qed
13.3. Functions

Remark 965. It is not straightforward to formalize the notion of correspondence between two values. We will synonymously use the terms mapping, function, transformation and operator. We will define functions as special binary relations in definition 968. Despite this being a standard practice, this has several drawbacks:

- There are some more mappings than the functions defined in definition 968. For example, assigning to a set $A$ its power set $P(A)$ can be cannot be regarded as a function because its domain and range should both be the set of all sets whose existence is inconsistent with ZFC by corollary 994 (a).

- We often work with spaces that have some additional structure in addition to being sets. In this case, we are often only interested in maps that preserve this structure. This is the case with group homomorphisms, for example.

In terms of first-order structures, not every function is a homomorphism. This is a motivating example for the benefits of category theory, where the notion of morphism is able to capture this additional structure.

It is often important to consider functions that are not homomorphisms, however. For example, function spaces over $\mathbb{R}$ contain some very complicated functions that are not field homomorphisms, order homomorphisms nor continuous functions and thus do not aim to preserve structural properties of $\mathbb{R}$.

- Several generalizations of the standard notion of a function are often used. These include multi-valued and partial functions. Both are not functions, strictly speaking.

For simplicity of exposition, we take multi-valued functions as primitive notions and define single-valued functions as special cases. This is actually done in [Aut20, def. 2.31] and [Kel55, p. 8] except that the corresponding authors conflate multi-valued functions and relations.

- Set-theoretic functions are often used in contexts where they do not refer to the intuitive notion of a mapping. Examples include Cartesian products and indexed families.

Definition 966. A multi-valued function from $A$ to $B$ is simply a binary relation $(F, A, B)$. For us the difference between a multi-valued function and a relation is merely in how we treat them. Multi-valued function are also called set-valued.

As discussed in definition 953 (a), the common practice of identifying a multi-valued function $(F, A, B)$ with its graph $F$ with no regard to its signature $(A, B)$ has serious drawbacks that we wish to avoid.

We use the more established notation $F : A \rightharpoonup B$ rather than $(F, A, B)$ and call the string of symbols $A \rightharpoonup B$ the signature of $F$ rather than the pair $(A, B)$.

(a) The value of $F$ at $x$ is

$$F(x) := \{y \in B \mid (x, y) \in F\}.$$
In case $x$ is not a concrete value, then $F(x)$ stands for the function $F$ itself. In other words, $F(x)$ refers to a member of $B$ if $x$ is a bound variable and $F(x)$ refers to the function $F$ if $x$ is a free variable.

(b) We also define the value of $F$ at a subset $X$ of $A$ as

$$F[X] := \bigcup \{F(x) \mid x \in X\}.$$  

This is also called the action of $F$ on the set $X$ or the image of $X$ under $F$. We also refer to $F(x)$ as the image of $x$ under $F$ because $F(x) = F[\{x\}]$.

The notation $f[X]$ is mostly used with set theory, for example in [Aut20, def. 2.31] and [End77, p. 44], however the following notation is usually used preferred of set theory:

$$f(X) = \bigcup \{F(x) \mid x \in X\}.$$  

We use $f[X]$ almost exclusively for section 13 (Set theory) because of the following nasty ambiguity: If $A$ is a transitive set, then $x \in A$ implies that $x \subseteq A$ yet $f(x) \neq f[\{x\}]$. For example, if $A = \{\emptyset, \{\emptyset\}\}$, then

$$f[\{\emptyset\}] = \{f(\emptyset)\} \neq f(\{\emptyset\}).$$  

This may be very confusing when dealing with ordinals.

(c) As mentioned in definition 966 (a), given a function $f : A \to B$, we sometimes use the notation $f(x)$ where $x$ is a free variable in the sense of definition 823 (j).

If $A = A_1 \times \cdots \times A_n$ is a finite Cartesian product, we instead use the notation $f(x_1, \ldots, x_n)$ and regard $x_1, \ldots, x_n$ as free variables that have no assigned value.

The variables are called arguments or sometimes parameters, although the latter term is a bit overloaded. This notion is somewhat informal and depends on the context since $A$ can usually be represented as a Cartesian product in different ways and with different arities. For example, if $A = B \times C$, we can write both $f(a)$ and $f(b, c)$ and the function has a different number of parameters in each case. In practice the number of arguments is usually clear from the context. We sometimes use $\vec{a}$ when we regard $a$ as a tuple.

For example, in classical first-order semantics, to each $n$-ary functional symbol there corresponds an $n$-ary function with the unambiguous signature $X^n \to X$.

When working over a vector space like $\mathbb{R}^2$, on the other side, depending on the context we regard functionals as either unary or binary functions.

We sometimes refer to $f$ as a dependent variable since it depends on its arguments. In this later case, we call the arguments independent variables.

Common terms for functions with low arity:

- **Nullary** for $n = 0$.
- **Unary** or **univariate** for $n = 1$.
- **Binary** or **bivariate** for $n = 2$.
- **Ternary** for $n = 3$.
- **Multivariate** for $n > 1$.

The following terminology is consistent with relations:

(d) The **graph** $\text{gph}(F)$ of $F$ is the graph of the relation $F$. This is consistent with definition 953 (a).

(e) The **restriction** of $F : A \rightarrow B$ to $X \subseteq A$ is the multi-valued function $F|_X : X \rightarrow B$. We say that $F$ is an **extension** of $F|_X$. This is consistent with definition 954 (c).

[La20a] Dually, if we restrict $B$ to its subset $Y$, we say that the obtained function is a **corestriction**.

(f) The **composition** $G \circ F$ of two multi-valued functions $F : A \rightarrow B$ and $G : B \rightarrow C$ is the function

$$[G \circ F](x) := G(F(x)).$$

The square brackets around $G \circ F$ are not a special notation, but rather another pair of delimiters that looks different from parentheses for the sake of reducing visual clutter. This definition is consistent with definition 954 (e).

(g) A function $F : A \rightarrow B$ is **empty** if $\text{gph}(F) = \emptyset$. This is consistent with definition 954 (a).

Note that if $F$ is a total empty multi-valued function, then $A = \emptyset$ because otherwise $F$ would itself be nonempty.

The following terminology is inconsistent with relations:

(h) The **arity** of a multi-valued function is its number of arguments. This is not to be confused with relation arity — functions are always binary relations.

(i) The term **total multi-valued function** means that $\text{dom}(F) = A$, i.e. that $F(x) \neq \emptyset$ for all $x \in A$. This is very different from total binary relations as defined in definition 954 (k).

(j) The **inverse** $F^{-1} : B \rightarrow A$ of a multi-valued function $F : A \rightarrow B$ is the multi-valued function in which assigns to every element $y$ of the image $\text{img}(F)$ the set of all $x \in A$ such that $y \in F(x)$. It is therefore the converse binary relation of $F$.

[kLei16] exer. 3.1.1

(k) The function corresponding to the diagonal relation $\Delta A$ as defined in definition 954 (d) is called the **identity function** on $A$ and denoted by $\text{id}_A$. The **diagonal function** is instead defined as $f : A \rightarrow A^2$ is defined as $f(x) := (x, x)$. This is mostly used in category theory.
We define some additional terminology:

(l) The **image** \( \text{img}(F) \) of \( F \) is the set of all \( y \in B \) that belong to the set \( F(x) \) for at least one \( x \in A \). It is the same as the value \( F(A) \).

(m) The **domain** \( \text{dom}(F) \) of \( F \) is the set of all values \( x \) for which \( F(x) \neq \emptyset \). When regarding \( F \) as a relation, the domain can be defined as the set 

\[
\text{dom}(R) := \{ x \in A \mid \exists y \in B. (x, y) \in R \}.
\]

(n) The **range** \( \text{range}(F) \) is simply the set \( B \). It is also called the **codomain** of \( F \).

(o) Although the terms “composition” and “superposition” are used interchangeably, for example in [Fic68a, No25], the term “superposition” often refers to a certain generalization of **function composition**. The latter usage can be found in [Yab86, p. 16].

If we are given the family of functions \( \{F_k : A \to B_k\}_{k=1}^n \) and the function \( G : B_1 \times \cdots \times B_n \to C \), we defined their **superposition** is

\[
H : A \to C \quad H(x) := G(F_1(x), \ldots, F_n(x)).
\]

(p) Functions from a set to itself (e.g. \( F : A \to A \)) are called **endofunctions**.

(q) For a fixed set \( Y \subseteq B \), its **large preimage** or simply **preimage** under \( F : A \to B \) is the image of \( Y \) under the inverse function \( F^{-1} : B \to A \). For a single value \( y \in B \), we call \( F^{-1}(y) \) the **fiber** of \( y \) under \( F \).

(r) Analogously, we define its **small preimage** as

\[
F^{-1}[Y] := \{ x \in A : F(x) \subseteq Y \}.
\]

Proposition 967. **Multi-valued functions** have the following basic properties:

(a) **Composition** is associative. That is, for any three functions \( F : A \to B, G : B \to C \) and \( H : C \to D \) we have 

\[
H \circ (G \circ F) = (H \circ G) \circ F.
\]

We will henceforth simply write \( H \circ G \circ F \).

(b) If \( F : A \to B \) and \( G : B \to C \) are multi-valued functions, then

\[
(G \circ F)^{-1} = F^{-1} \circ G^{-1}.
\]

Proof.

**Proof of 967 (a).** Let \( a \in A \). Then in order for \( d \in D \) to belong to \( [(H \circ G) \circ F](a) \), there must exist values \( b \in B \) and \( c \in C \) such that \( b \in F(a) \) and \( c \in G(b) \) and \( d \in H(c) \). Clearly this is also the condition for \( d \) to belong to \( [H \circ (G \circ F)](a) \).
Proof of 967 (b). Since $G \circ F$ has signature $A \to C$, clearly $[G \circ F]^{-1}$ has signature $C \to A$. Let $c \in C$.

- If $[G \circ F]^{-1}(c)$ is empty, $c \notin \text{img}(G \circ F)$, hence either $G^{-1}(c)$ is empty or is nonempty, but disjoint from $\text{img}(F)$. Hence, $[F^{-1} \circ G^{-1}](c)$ is also empty.
- Suppose that $[G \circ F]^{-1}(c)$ is not empty and let $a \in [G \circ F]^{-1}(c)$.
  By definition, there exists $b \in B$ and such that $b \in F(a)$ and $c \in G(b)$. Hence, $b \in G^{-1}(c)$ and $a \in F^{-1}(b)$, which implies that the image of $c$ under the composition $F^{-1} \circ G^{-1}$ also contains $a$.

In both cases, for every $c \in C$ we have

$$[G \circ F]^{-1}(c) = [F^{-1} \circ G^{-1}](c).$$

Hence, the two multi-valued functions are equal. \hfill\Box

Definition 968. Although multi-valued functions are very general, they are not nearly as useful nor are not studied nearly as extensively as single-valued functions.

The multi-valued function $F : A \rightrightarrows B$ is called a single-valued function if $F(x)$ is a singleton set for each $x \in A$. In this case, we write $F : A \to B$ rather than $F : A \rightrightarrows B$.

All terminology from definition 966 holds for single-valued functions.

By convention, when both single-valued and multi-valued functions are involved, the former are denoted using lowercase letters and the latter using uppercase letters.

Strictly speaking, the value $f(x)$ of a single-valued function $f : A \to B$ is a singleton set. It is common practice (e.g. in [Aut20, def. 3.1] and [Kel55, p. 10]) to define the value of a single-valued function to be an element of $B$ rather than a subset of $B$. Unless this would be confusing, we identify $f(x)$ with its only element due to the convention established in remark 930.

More precisely, single-valued functions satisfy the following predicate formula, which states that the free variable $\chi$ is a function from $\alpha$ to $\beta$:

$$\text{IsFun}[\chi, \alpha, \beta] := \forall \xi \in \alpha . \exists! \eta \in \beta . \exists \zeta \in \chi . \text{IsPair}[\zeta, \xi, \eta].$$

Unless otherwise noted, we will now conflate the terms “function” and “single-valued function”.

(a) We can simplify definition 966 (b) drastically:

$$f[A] = \bigcup \{f(x) \mid x \in A\} = \{f(x) \mid x \in A\}.$$

As mentioned in definition 966 (b), we usually prefer the notation $f[A]$ outside section 13 (Set theory) where we are less prone to ambiguity.
(b) Given the two-argument function \( f : A \times B \to C \), we can define another function \( g : A \to \text{fun}(B, C) \) as
\[
g(x) := (y \mapsto f(x, y)).
\]
This process is called \textbf{currying} after Haskell Curry. Obtaining \( f \) from \( g \) is instead called \textbf{uncurrying}.

Currying is useful if we have somehow fixed a value \( x_0 \in A \), in which case we can “get rid” of one argument by introducing some shortcut for the function \( g(x_0) : B \to C \) for the sake of reducing notational clutter. See definition 204 (a) and the proof of proposition 1065 for example of how this is useful in the wild. See proposition 1166 for a categorical description of currying.

(c) If \( f : A \to B \) is a single-valued function, \( F : A \rightrightarrows B \) is a multi-valued function and \( \text{gph}(f) \subseteq \text{gph}(F) \), we say that \( f \) is a \textbf{selection} of \( F \).

(d) We denote the set of all single-valued total functions from \( A \) to \( B \) by \( \text{fun}(A, B) \). Other accepted notation is either \( \text{Set}(A, B) \) (which is consistent with morphisms in the category of sets) or by \( B^A \) (which is consistent with cardinal exponentiation). We abbreviate \( \text{fun}(A, A) \) as \( \text{fun}(A) \).

\textbf{Proposition 969.} \textit{Single-valued functions have the following properties when regarded as multi-valued functions:}

(a) Single-valued functions are \textbf{total} as multi-valued functions. As a consequence, a multi-valued function cannot have a selection unless it is \textbf{total}.

This explains why there is no established terminology for the set \( A \) analogous to “range” for \( B \) — we rarely need to differentiate between \( A \) and \( \text{dom}(f) \).

(b) The \textbf{composition} of single-valued functions is a single-valued function.

(c) The \textbf{large preimages} and \textbf{small preimages} of single-valued functions are identical. We restrict ourselves to large preimages with the notation \( f^{-1}(x) \) and refer to them simply as preimages.

Note that preimages are multi-valued in general.

\textbf{Proof.} Trivial. \hfill \square

\textbf{Definition 970.} A \textbf{multi-valued function} that is otherwise single-valued, but not necessarily \textbf{total} is called a \textbf{partial function}. That is, \( f : A \to B \) is a partial function if \( f(x) \) has at most one element for every \( x \in A \).

\textbf{Definition 971.} We introduce the following terminology for invertibility of a (single-valued) function \( f : A \to B \):
(a) We say that $f$ is **injective** or **one-to-one** if any of the following equivalent conditions hold:

(i) For any $y \in B$ there exists at most one $x \in A$ such that $f(x) = y$.
That is, each point in $B$ is the image of at most one point in $A$.

(ii) For all $x_1, x_2 \in A$, the equality $f(x_1) = f(x_2)$ implies $x_1 = x_2$.
The contrapositive of this statement is that different points in $A$ have different images under $f$.

(iii) The inverse is a partial single-valued function.

(b) $f$ is called **surjective** or **onto** if any of the following equivalent conditions hold:

(i) For any $y \in B$ there exists at least one $x \in A$ such that $f(x) = y$.
That is, each point in $B$ is the image of at least one point in $A$. Hence, the image of $f$ is the entire range $B$.

(ii) For all $y_1, y_2 \in B$, the equality $f^{-1}(y_1) = f^{-1}(y_2)$ implies $y_1 = y_2$.
Without surjectivity, the above holds only for the points in the image of $f$.

(iii) The inverse is a total multi-valued function.

(c) Finally, $f$ is called **bijective** if any of the following equivalent conditions hold:

(i) It is both injective and surjective.

(ii) For any $y \in B$ there exists exactly one $x \in A$ such that $f(x) = y$.

(iii) The inverse is a total single-valued function.

**Proof of correctness.** The equivalences are trivial to verify.

**Proposition 972.** The composition of injective (resp. surjective or bijective) functions is injective (resp. surjective or bijective).

Compare this result to proposition 1128 (i) and proposition 974.

**Proof.** Let $f : A \to B$ and $g : B \to C$ be arbitrary functions and define $h := g \circ f : A \to C$.

**Proof of injectivity.** Suppose that $f$ and $g$ are injective. We will use definition 971 (a ii).

Let $x_1, x_2 \in A$ and suppose that $h(x_1) = h(x_2)$, that is, $g(f(x_1)) = g(f(x_2))$. Then $f(x_1) = f(x_2)$ since $g$ is injective and $x_1 = x_2$ since $f$ is injective.

Since $x_1$ and $x_2$ were arbitrary, we conclude that $h$ is also injective.

**Proof of surjectivity.** Suppose that $f$ and $g$ are surjective. We will use definition 971 (a i).

Let $z \in C$. Then there exists some $y \in B$ such that $g(y) = z$ because $g$ is surjective and similarly there exists some $x \in B$ such that $f(x) = y$ because $f$ is surjective. Thus, $h(x) = g(f(x)) = z$.

Since $z$ was arbitrary, we conclude that $h$ is also surjective.

**Proof of bijectivity.** We have shown that if $f$ and $g$ are both injective or surjective, then, so is $h$. Hence, if $f$ and $g$ are bijective, so is $h$. \[\square\]
Lemma 973. Let \( \{f_k\}_{k \in \mathcal{K}} \) be an indexed family of functions where \( f_k : A \to B_k \) in \( k \in \mathcal{K} \). Then if at least one of the functions is injective, the diagonal product
\[
f : A \to \prod_{k \in \mathcal{K}} B_k
\]
\[
f(x) := \{f_k(x)\}_{k \in \mathcal{K}}
\]
is also injective.

Proof. Suppose that \( f_{k_0} \) is injective. Let \( f(x_1) = f(x_2) \). Then \( \{f_k(x_1)\}_{k \in \mathcal{K}} = \{f_k(x_2)\}_{k \in \mathcal{K}} \) and thus \( f_{k_0}(x) = f_{k_0}(y) \). Since \( f_{k_0} \) is injective, we conclude that \( x_1 = x_2 \). Since \( x_1 \) and \( x_2 \) were arbitrary, we conclude that \( f \) is also injective.

Proposition 974. The superposition of injective functions is injective.

Compare this result to proposition 972.

Proof. Suppose that we are given injective functions \( f_k : A \nrightarrow B_k, k = 1, ..., n \) and \( g : B_1 \times \cdots \times B_n \nrightarrow C \). From lemma 973 it follows that the function
\[
d : A \to B_1 \times \cdots \times B_n d(x) := (f_1(x), ..., f_n(x))
\]
is injective.

Then from proposition 972 it follows that the desired superposition
\[
h : A \to C \quad h(x) := g(f_1(x), ..., f_n(x)) = g(d(x))
\]
is injective.

Remark 975. We can represent a multi-valued functions \( F : A \nrightarrow B \) as the indexed family \( \{F(a)\}_{a \in A} \). This indexed family is itself a single-valued function \( G \) from \( A \) to \( \text{pow}(B) \). This is an alternative to our approach to define multi-valued functions as a basic notion. The latter approach is used, for example, in [Phe93, def. 2.3]. The pair \( (x, y) \in A \times B \) belongs to the relation \( F \) if and only if there exists a subset \( Y \subseteq B \) such that \( y \in Y \) and the pair \( (x, Y) \) belongs to the relation \( G \).

The downside of the latter approach is that notions such as the image, range and inverse of the multi-valued function have a very different and much less useful meaning and notions such as endofunctions cannot even be defined.

Proposition 976. For any function \( f : A \to B \) we have
\[
(a) \text{ If } X \subseteq A, \text{ then } X \subseteq f^{-1}[f[X]] \text{ with equality holding if } f \text{ is injective.}
\]
\[
(b) \text{ If } Y \subseteq B, \text{ then } f[f^{-1}[Y]] \subseteq Y \text{ with equality holding if } f \text{ is surjective.}
\]

Proof.
**Proof of 976 (a).** If \( x \in X \), clearly \( x \in f^{-1}(f(x)) \). Thus,

\[
X \subseteq f^{-1}[f[X]].
\]

Now suppose that \( f \) is injective and let \( x \in f^{-1}[f[X]] \). There exists some \( y \in f[X] \) such that \( x \in f^{-1}(y) \) and some \( z \in X \) such that \( y = f(z) \). Since \( f \) is injective and \( f(x) = y = f(z) \), it follows that \( x = z \) and thus \( x \in X \). Since \( x \) was chosen arbitrarily from \( f^{-1}[f[X]] \), we conclude that

\[
f^{-1}[f[X]] \subseteq X.
\]

**Proof of 976 (b).** If \( y \in f[f^{-1}[Y]] \), there exists some \( x \in f^{-1}[Y] \) such that \( f(x) = y \). Furthermore, there also exists some \( t \in Y \) such that \( x \in f^{-1}(t) \). Hence, \( y = f(x) = t \) and \( y \in Y \). Therefore,

\[
f[f^{-1}[Y]] \subseteq Y.
\]

Now suppose that \( f \) is surjective and let \( y \in Y \). Then from surjectivity it follows that there exists some \( x \in X \) such that \( f(x) = y \). Hence, \( x \in f^{-1}(y) \) and \( y = f(x) \in f[f^{-1}(y)] \). Since \( y \) was chosen arbitrarily from \( Y \), we conclude that

\[
Y \subseteq f[f^{-1}[Y]].
\]

\[\square\]

**Proposition 977.** Images of sets under \( f : A \to B \) have the following basic properties:

(a) If \( A_1 \subseteq A_2 \), then \( f[A_1] \subseteq f[A_2] \).

(b) For any indexed family \( \{A_k\}_{k \in \mathcal{X}} \subseteq A \) of subsets of \( A \) we have the equality

\[
f \left[ \bigcup_{k \in \mathcal{X}} A_k \right] = \bigcup_{k \in \mathcal{X}} f[A_k]. \tag{339}\]

(c) For any indexed family \( \{A_k\}_{k \in \mathcal{X}} \) of subsets of \( A \) we have the inclusion

\[
f \left[ \bigcap_{k \in \mathcal{X}} A_k \right] \subseteq \bigcap_{k \in \mathcal{X}} f[A_k]. \tag{340}\]

Equality in (340) holds if \( f \) is injective. If \( f \) is not injective, for example if both \( A \) and \( B \) are nonempty, \( A_1 \) and \( A_2 \) are disjoint subsets of \( A \) and \( f[A_1] = f[A_2] = B \), then

\[f[A_1 \cap A_2] = f[\emptyset] = \emptyset \subsetneq f[A_1] \cap f[A_2] = B.\]

(d) For any two subsets \( A_1 \) and \( A_2 \) of \( A \) we have the inclusion

\[
f[A_1] \setminus f[A_2] \subseteq f[A_1 \setminus A_2]. \tag{341}\]

Equality in (341) holds if \( f \) is injective. If \( f \) is not injective, for example if \( A_1 \subsetneq A_2 \), but \( f[A_1] = f[A_2] \), then

\[f[A_1] \setminus f[A_2] = \emptyset \subsetneq f[A_1 \setminus A_2].\]
Compare this result to the more well-behaved properties of preimages described in proposition 978.

Proof.

**Proof of 977 (a).** If \( x \in A_1 \), then \( x \in A_2 \) and hence \( f(x) \in f[A_2] \). Therefore, \( f[A_1] \subseteq f[A_2] \).

**Proof of 977 (b).** If \( x_0 \in A_{k_0} \) for some \( k_0 \in \mathcal{K} \), clearly
\[
f(x_0) \in f[A_{k_0}] \subseteq \bigcup_{k \in \mathcal{K}} f[A_k].
\]
Therefore,
\[
f \left[ \bigcup_{k \in \mathcal{K}} A_k \right] \subseteq \bigcup_{k \in \mathcal{K}} f[A_k].
\]
Conversely, if \( y_0 \in f[A_{k_0}] \) for some \( k_0 \in \mathcal{K} \), by proposition 977 (a) obviously
\[
y_0 \in f \left[ \bigcup_{k \in \mathcal{K}} A_k \right].
\]
Therefore,
\[
f \left[ \bigcup_{k \in \mathcal{K}} A_k \right] \supseteq \bigcup_{k \in \mathcal{K}} f[A_k].
\]
Hence, (339) holds.

**Proof of 977 (c).** If \( x_0 \) belongs to \( \bigcap_{k \in \mathcal{K}} A_k \), then \( x_0 \) belongs to \( A_k \) for all \( k \in \mathcal{K} \). It follows that \( f(x_0) \) belongs to \( f[A_k] \) for all \( k \in \mathcal{K} \) and hence to their intersection. Therefore, the inclusion (340) holds.

Now suppose that \( f \) is injective. Let \( y_0 \) be a point in the intersection \( \bigcap_{k \in \mathcal{K}} f[A_k] \). We thus have \( y_0 \in f[A_k] \) for all \( k \in \mathcal{K} \). Since \( f \) is injective, for each \( k \in \mathcal{K} \) there exists a unique \( x_k \in A_k \) such that \( f(x_k) = y_0 \). Again because of injectivity of \( f \), all these elements are equal because \( f(x_k) = f(x_m) = y_0 \) for \( k, m \in \mathcal{K} \). Hence,
\[
y_0 \in f \left[ \bigcap_{k \in \mathcal{K}} A_k \right].
\]
Therefore, the reverse inclusion in (340) holds if \( f \) is injective.

**Proof of 977 (d).** If \( f[A_1] \setminus f[A_2] \) is empty, (341) obviously holds. Suppose that it is nonempty and let \( y_0 \in f[A_1] \setminus f[A_2] \).

Then there exists a point \( x_0 \in A_1 \) such that \( f(x_0) = y_0 \). It cannot be that \( x_0 \in A_2 \) because otherwise \( y_0 = f(x_0) \in f[A_2] \), which would contradict our choice of \( y_0 \). Hence, \( x_0 \in A_1 \setminus A_2 \) and \( y_0 \in f(A_1 \setminus A_2) \).

Since \( y_0 \) was chosen arbitrarily, we conclude that the inclusion (341) holds.
Conversely, suppose that $f$ is injective. If $f(A_1 \setminus A_2)$ is empty, by (341) the set $f[A_1] \setminus f[A_2]$ is also empty and the converse holds.

Now suppose that it is nonempty and let $y_0 \in f(A_1 \setminus A_2)$. Then there exists a point $x_0 \in A_1 \setminus A_2$ such that $f(x_0) = y_0$. Furthermore, since $f$ is injective, $x_0$ is the only preimage of $y_0$ and hence $f(x_0) \in f[A_1] \setminus f[A_2]$, which proves the reverse inclusion in (341).

**Proposition 978.** Function preimages have the following basic properties:

(a) If $B_1 \subseteq B_2$, then $f^{-1}[B_1] \subseteq f^{-1}[B_2]$.

(b) For any indexed family $\{B_k\}_{k \in \mathcal{K}} \subseteq B$ of subsets of $B$ we have the equality

$$f^{-1}\left[\bigcup_{k \in \mathcal{K}} B_k\right] = \bigcup_{k \in \mathcal{K}} f^{-1}[B_k].$$  \tag{342}

(c) For any indexed family $\{B_k\}_{k \in \mathcal{K}}$ of subsets of $B$ we have the equality

$$f^{-1}\left[\bigcap_{k \in \mathcal{K}} B_k\right] = \bigcap_{k \in \mathcal{K}} f^{-1}[B_k].$$  \tag{343}

(d) For any two subsets $B_1$ and $B_2$ of $B$ we have the equality

$$f^{-1}[B_1] \setminus f^{-1}[B_2] = f^{-1}[B_1 \setminus B_2].$$  \tag{344}

Compare this result to the less well-behaved properties of images described in proposition 977.

**Proof.**

**Proof of 977 (a).** Analogous to proposition 977 (a).

**Proof of 977 (b).** Analogous to proposition 977 (b).

**Proof of 977 (c).** If $y_0$ belongs to $\bigcap_{k \in \mathcal{K}} B_k$, then $y_0$ belongs to $B_k$ for all $k \in \mathcal{K}$. It follows that $f(y_0) \subseteq f^{-1}[B_k]$ for all $k \in \mathcal{K}$ and hence it is also a subset of their intersection. Therefore,

$$f^{-1}\left[\bigcap_{k \in \mathcal{K}} B_k\right] \subseteq \bigcap_{k \in \mathcal{K}} f^{-1}[B_k].$$

Conversely, if $x_0 \in \bigcap_{k \in \mathcal{K}} f^{-1}[B_k]$, it belongs to $f^{-1}[B_k]$ for all $k \in \mathcal{K}$. Clearly then $f(x_0) \in B_k$ for all $k \in \mathcal{K}$ and thus $f(x_0) \in \bigcap_{k \in \mathcal{K}} B_k$. Hence, by proposition 978 (a) we have

$$f^{-1}(f(x_0)) \subseteq f^{-1}\left[\bigcap_{k \in \mathcal{K}} B_k\right].$$

Since $x_0 \in f^{-1}(f(x_0))$,

$$x_0 \in f^{-1}\left[\bigcap_{k \in \mathcal{K}} B_k\right].$$
Since $x_0$ was chosen arbitrarily from $\bigcap_{k \in \mathcal{K}} f^{-1}[B_k]$, we can conclude that

$$\bigcap_{k \in \mathcal{K}} f^{-1}[B_k] \in f^{-1}\left[\bigcap_{k \in \mathcal{K}} B_k\right].$$

Hence, (343) holds.

**Proof of 978 (d).** If $y_0 \in B_1 \setminus B_2$, there exists a point $x_1 \in B_1$ such that $f(x_1) = y_0$. Aiming at a contradiction, suppose that there exists a point $x_2 \in f^{-1}[B_2]$ such that $f(x_2) = y_0$. Then $y_0 = f(x_1) = f(x_2)$ implies that $f^{-1}(y_0) \subseteq f^{-1}[B_1] \cap f^{-1}[B_2]$. Proposition 978 (c) then in turn implies that $f^{-1}(y_0) \subseteq f^{-1}[B_1 \cap B_2]$ and hence by proposition 977 (a)

$$y_0 = f(f^{-1}(y_0)) \in f[f^{-1}[B_1 \cap B_2]] = B_1 \cap B_2,$$

which contradicts our choice of $y_0$. Since the choice of $y_0 \in B_1 \setminus B_2$, $x_1 \in f^{-1}(y_0) \cap B_1$ and $x_2 \in f^{-1}(y_0) \cap B_2$ was arbitrary, the obtained contradiction shows that

$$f^{-1}(B_1 \setminus B_2) \subseteq f^{-1}[B_1] \setminus f^{-1}[B_2].$$

Conversely, we have

$$f(f^{-1}[B_1 \setminus f^{-1}[B_2]]) \subseteq f( f^{-1}(B_1 \setminus B_2)) \subseteq B_1 \setminus B_2.$$ 

Hence,

$$f^{-1}[B_1 \setminus f^{-1}[B_2]] \subseteq f^{-1}\left(f\left( f^{-1}[B_1] \setminus f^{-1}[B_2]\right)\right).$$

Theorem 979 (Recursion theorem). Let $A$ be a nonempty set. Suppose that we are given some member $a_0$ of $A$ and some transformation $T : A \to A$. Then there exists a sequence $f : \omega \to A$ such that

- $f(n) = a_0$.
- For every $n \in \omega$ we have $f(succ(n)) = T(f(n))$.

Note that we do not yet use the notation $n + 1$ because we will use this theorem to define addition in the first place.

See remark 982 for a simpler and more conventional notation for recursion on $\omega$.

**Proof.** Let $G \subseteq \text{pow}(\omega \times A)$ be the set of all partial single-valued functions $g : \omega \to A$ such that

- If $g$ is defined at $\emptyset$, then $g(\emptyset) = a_0$.
- For every $n \in \omega$, if $g$ is defined at $\text{succ}(n)$, then $g$ is also defined at $n$ and $g(\text{succ}(n)) = T(f(n))$. 

493
Clearly $G$ is nonempty because the function $\{ (\emptyset, a_0) \}$ belongs to $G$.

The conditions imposed on the functions in $G$ ensure that every function is defined in some initial segment of the natural numbers. A more obvious approach is to require $g$ to be defined at $\text{succ}(n)$ if it is defined at $n$, however we are trying to prove that such a function exists in the first place.

Define $f := \bigcup G$. At this point $f$ is a multi-valued function. We must now show that $f$ has all the properties that we want.

**Proof of totality.** First, we will use theorem 943 (Induction via inductive sets) to show that $f$ is total. Clearly $\emptyset \in \text{dom } f$.

Now fix $n \in \text{dom } f$. Then there exists a function $g \in G$ defined at $n$.

- If $g$ is also defined at $\text{succ}(n)$, this directly proves that $\text{succ}(n) \in \text{dom } f$.
- If $g$ is not defined at $\text{succ}(n)$, consider
  \[ \hat{g} := g \cup \{(\text{succ}(n), T(g(n))\} \].

The function $\hat{g}$ is again a single-valued partial function and thus it belongs to $G$, hence $\text{succ}(n) \in \text{dom } f$.

Therefore, theorem 943 (Induction via inductive sets) allows us to conclude $f : \omega \rightarrow A$ is a total multi-valued function.

**Proof of single-valuedness.** Now that we know that $f$ is total, we will prove that it is single-valued and thus a function in the usual sense of the term.

Clearly $f$ is single-valued at $\emptyset$.

Now suppose that $f$ is single-valued at $n$. Since $f$ is total, there exist at least one partial function in $G$ that is defined at $\text{succ}(n)$, from which it follows that it is also defined at $n$. Let $g$ and $h$ both be such (single-valued partial) functions.

Then
\[ g(\text{succ}(n)) = T(g(n)) = T(f(n)) = T(h(n)) = h(\text{succ}(n)), \]

hence $g$ and $h$ coincide at $\text{succ}(n)$, which in turn implies that $f$ is single-valued at $\text{succ}(n)$.

Therefore, theorem 943 (Induction via inductive sets) allows us to conclude that $f$ is a single-valued total function.

**Proof of uniqueness.** Now that it is clear that $f$ satisfies the theorem, we must verify that it is unique.

Suppose that $f_1$ and $f_2$ both satisfy the theorem. Clearly $\emptyset \in H$. Fix some $n \neq \emptyset$ and suppose that $f_1(n) = f_2(n)$. Then
\[ f_2(\text{succ}(n)) = T(f_1(n)) = T(f_2(n)) = f_2(\text{succ}(n)). \]

Therefore, theorem 943 (Induction via inductive sets) allows us to conclude that $f_1 = f_2$. So there is at most one function that satisfies the theorem and we have already shown that $f$ is such a function. 

\[ \square \]
Definition 980. We will use theorem 979 (Recursion theorem) for defining arithmetic operations for natural numbers. There constructions will be more elaborate than the basic recursive sequences defined in e.g. theorem 390 (Banach’s fixed point theorem).

(a) We will represent the addition operation \( \oplus \) as follows: fix the first summand \( n \) and then define a function \( \oplus_n : \omega \rightarrow \omega \) such that \( k = \oplus_n(m) \) gives us \( n \oplus m = k \). This is a particular instance of currying.

Fix \( n \in \omega \), let \( A = \omega \) and define

\[
T_n : \omega \rightarrow \omega \\
T_n(k) := \text{succ}(k).
\]

Now we use the initial point \( a_n = n \) to construct the function \( \oplus_n \).

Define the addition function \( \oplus : \omega \times \omega \rightarrow \omega \) via its graph

\[
\{(n, m), \oplus_n(m) \mid n, m \in \omega \}.
\]

(b) We define natural number multiplication analogously. For each \( n \in \omega \), define \( \odot_n \) via

\[
T_n : \omega \rightarrow \omega \\
T_n(k) := k \oplus n.
\]

and \( a_n = 0 \) and then define \( \odot : \omega \times \omega \rightarrow \omega \) via its graph

\[
\{(n, m), \odot_n(m) \mid n, m \in \omega \}.
\]

Theorem 981. The smallest inductive set \( \omega \) is a model of Peano arithmetic with the following interpretation:

(a) Zero is interpreted as \( \emptyset \).

(b) The successor operation \( s \) is interpreted as \( \text{succ} \).

(c) Addition is interpreted by the \( \oplus \) function given in definition 980 (a).

(d) Similarly, multiplication is interpreted by \( \odot \) from definition 980 (b).

This is an extension of theorem 948.

Proof. We have already shown in theorem 948 that \( \omega \) satisfies the axioms (PA1)-(PA3).

Proof of PA4. For each \( n \in \omega \) the starting condition (i.e. \( m = \emptyset \)) in definition 980 (a) implies that \( n \oplus \emptyset = n \).

Proof of PA5. For each \( n \in \omega \) the transformation \( T_n \) in definition 980 (a) is defined, so that if \( \oplus_n(m) = k \), then \( \oplus_n(\text{succ}(m)) = \text{succ}(k) \).

It follows that \( n \oplus \text{succ}(m) = \text{succ}(n \oplus m) \) for all \( n, m \in \omega \).
Proof of PA6. Analogously to $\oplus$, the starting condition in definition 980 (b) implies that $n \odot \emptyset = \emptyset$ for every $n \in \omega$.

Proof of PA7. Analogously to $\oplus$, for each $n \in \omega$ the transformation $T_n$ in definition 980 (b) is defined, so that if $\odot_n(m) = k$, then $\odot_n(\text{succ}(m)) = k \oplus n$.

It follows that $n \odot \text{succ}(m) = n \odot m + n$ for all $n, m \in \omega$. 

Remark 982. With the availability of natural numbers, instead of the tedious constructions in definition 980 (a), we can use a more conventional notation when applying theorem 979 (Recursion theorem).

As an example, we can define the Fibonacci sequence. The sequence is motivated by the problem in example 1280. In the notation of theorem 979 (Recursion theorem), we can define the Fibonacci sequence as follows:

$$T : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$$

$$T(a, b) := (b, a + b)$$

and

$$a_0 := (0, 1).$$

The recursion theorem gives us a sequence of pairs

$$\begin{align*}
(0, 1), (1, 1), (1, 2), (2, 3), (3, 5), (5, 8), (8, 13), \ldots \\
a_0, a_1, a_2, a_3, a_4, a_5, a_6
\end{align*}$$

The pairs are only a technicality because otherwise we would not be able to define the sequence $\{a_k\}_{k=1}^\infty$.

By taking the second element of each pair, we obtain the sequence

$$\begin{align*}
1, 1, 2, 3, 5, 8, 13, \ldots \\
b_0, b_1, b_2, b_3, b_4, b_5, b_6
\end{align*}$$

In order to obtain the Fibonacci sequence, we must prefix the sequence $\{b_k\}_{k=0}^\infty$ with 0. This is undoubtedly much more complicated than writing

$$b_k := \begin{cases} 
0, & k = 0 \\
1, & k = 1 \\
b_{k-1} + b_{k-2} & k > 1.
\end{cases}$$

To see that the latter notation is merely syntax sugar, note that the other sequence can be written as

$$a_k := \begin{cases} 
(0, 1), & k = 0 \\
T(a_{k-1}), & k > 1.
\end{cases}$$
13.4. Zermelo-Fraenkel set theory

**Definition 983.** Let \( \mathcal{A} \) be a (potentially empty) family of nonempty sets. A **choice function** on \( \mathcal{A} \) is a (total) function \( c : \mathcal{A} \to \bigcup \mathcal{A} \) such that \( c(A) \in A \) for all \( A \in \mathcal{A} \).

This is formally the same as a tuple defined in definition 950 (c).

A choice function on \( \mathcal{A} \) “chooses” an element out of each member of \( \mathcal{A} \). We sometimes have a canonical choice function, for example in proposition 642, however for general quotient sets the existence of a choice function is not by any means obvious.

The existence of a choice function for family of nonempty sets is an important axiom of ZFC — see definition 984 (g).

**Definition 984.** The first-order theory commonly abbreviated as ZFC is based on the same language as naïve set theory, but with different axioms. The three letters refer to:

- Zermelo, who formulated the entire theory except for the **axiom schema of replacement** and the **axiom of foundation**.

- Fraenkel, who simultaneously with Skolem reformulated the theory within first-order logic while also introducing the axiom schema of replacement.

- The **axiom of choice**, which is part of Zermelo’s original theory, but is controversial enough to attract special attention — see theorem 990 (Axiom of choice equivalences).

We are usually only interested in either ZFC, which include all axioms listed in this definition. If we wish to avoid the axiom of choice — for example when proving the equivalences in theorem 990 (Axiom of choice equivalences) — we instead use ZF, which excludes the axiom of choice. The abbreviation of the latter theory is inaccurate historically, but is nevertheless established.

If we wish to instead exclude the axiom of replacement, we obtain the theory Z, however without context it is unclear whether the axiom of choice is included in Z or not.

See proposition 985 for proofs of existence of common sets.

The full list of axioms is:

(a) The **axiom of extensionality**, as defined in definition 924 (a). This is also the only axiom of the theory that does not deal with existence.

(b) The **axiom schema of specification**, also known as the axiom schema of separation or of **restricted comprehension**, states that given a set \( A \), any formula defines a subset of \( A \). For each formula \( \varphi \) containing neither \( \tau \) nor \( \sigma \) as free variables, the following is an axiom:

\[
\forall \sigma. \exists \tau. \forall \xi. (\xi \in \tau \iff \varphi \land \xi \in \sigma).
\] (345)

As explained in definition 924 (b) and definition 928, this set may depend on parameters, which are other sets. We must formally take the universal closure of this set to quantify over all possible values for the parameters.

Compare this axiom to **unrestricted comprehension**. Informally, this axiom can be obtained by taking the result of unrestricted comprehension and intersecting it with
some set $A$. As mentioned in definition 928, in set-builder notation such a set is usually denoted by

$$\{x \in B \mid \phi[x, u_1, \ldots, u_n]\}.$$  

Unlike unrestricted comprehension some definable subset of the universe in the metatheory no longer have a corresponding set within the object logic.

(c) The **axiom of power sets** states that every set has a corresponding **power set**. Symbolically,

$$\forall \tau . \exists \sigma . \text{IsPowerSet}[\sigma, \tau]. \quad (346)$$

(d) The **axiom of unions** states that for every set there exists another set that is its **union**. Symbolically,

$$\forall \tau . \exists \sigma . \text{IsUnion}[\sigma, \tau]. \quad (347)$$

(e) The **axiom of pairing** states that for any sets $A$ and $B$ there exists another set that contains exactly $A$ and $B$. This is the set $\{A, B\}$ in set-builder notation. Symbolically,

$$\forall \tau . \forall \rho . \forall \xi . (\xi \in \rho \equiv (\xi \equiv \tau \lor \xi \equiv \sigma)). \quad (348)$$

(f) The **axiom of infinity** states that an **inductive set** exists. Symbolically,

$$\exists \tau . \text{IsInductive}[\tau]. \quad (349)$$

This axiom is a simple and convenient way to state that infinite sets exist. Without it, we can only deal with finite sets unless we include some other axiom to replace it.

(g) The **axiom of choice** states that a choice function exists for any family of nonempty sets. To state the axiom via a formula, we will avoid functions and only state it in terms of the image of the choice function. That is, we will formulate that for each family $\mathcal{A}$ of nonempty sets there exists a set $B$ such that $A \cap B$ is a singleton set for each $A \in \mathcal{A}$. Symbolically,

$$\forall \tau . \left( \forall \xi . \xi \in \tau \equiv \neg \text{IsEmpty}[[\xi]] \rightarrow (\exists \sigma . \forall \xi . \forall \eta . \exists! \xi . \eta \in \xi \land \eta \in \xi) \right) \quad (350)$$

where we have used the convention regarding existence and uniqueness described in remark 825 (d).

See theorem 990 (Axiom of choice equivalences) for more statements equivalent to this axiom.

(h) The **axiom schema of replacement** roughly states that every mapping that is definable via a formula of ZFC is a function. As we have done for the **axiom of choice**,
we only formulate the axiom via the image of the function. More concretely, given a formula \(\varphi\) not containing \(\tau\) nor \(\sigma\) as free variables, the following is an axiom:

\[
\forall \tau. \left( \forall \xi \in \tau. \exists! \eta. \varphi \rightarrow \left( \exists \sigma. \forall \eta. \left( \eta \in \sigma \rightarrow \exists \xi \in \tau. \varphi \right) \right) \right).
\]  

(351)

As is the case with the axiom schema of specification, the formula \(\varphi\) may depend on parameters, in which case we use its universal closure.

This axiom is useful in cases where it is impossible or at least difficult to construct a function, for example in proposition 985 (m) or theorem 1029 (Hartogs' lemma). This is the case, in general, when dealing with indexed families rather than functions.

This is the axiom that makes ZFC require large models — see theorem 1104.

(i) The axiom of foundation states that every nonempty set contains a member disjoint from the set itself. Symbolically,

\[
\forall \tau. \left( \neg \text{IsEmpty}[\tau] \rightarrow \exists \sigma \in \tau. \neg \exists \xi. \left( \xi \in \tau \land \xi \in \sigma \right) \right).
\]  

(352)

This is a very powerful axiom because it shows that set membership in ZFC is well-founded — see proposition 993. It is equivalent to theorem 1093 (Axiom of regularity) and is often itself called the axiom of regularity.

**Proposition 985.** We will now prove that all sets we have considered up until now in section 13 (Set theory) are sets in ZFC. The uniqueness in all cases follows from the axiom of extensionality.

A very fundamental existence result is provided by the fact that we are assuming standard and transitive models of ZFC. Let \(V = (V, I)\) be such a model. Then if \(v \in V\) and \(u \subseteq v\), transitivity implies that \(u \in V\). Since the model is also standard, this shows that both \(u\) and \(v\) are sets within the object theory. Thus, if \(A\) is a set within the object theory and if \(B \subseteq A\) within the metatheory, then necessarily \(B\) itself is a set within the object theory.

For example, proposition 985 (p) shows that the set \(\text{Un}(A, B)\) of functions exists within the object theory for any two sets \(A\) and \(B\) in the object theory. Therefore, every single function between \(A\) and \(B\) is a set within the object theory because it is a member of \(\text{Un}(A, B)\).

With that in mind, we will show the following:

(a) If \(A\) is a set, then for any formula \(\varphi\) the set \(\{ x \in A \mid \varphi[x] \}\) exists and, furthermore, it is a subset of \(A\). Only definable subsets of \(A\) can be described in this way, however. See proposition 985 (j).

(b) There exists a unique empty set, which we denote by \(\emptyset\).

(c) No universal set (set of all sets) exists.

(d) For every set \(A\), there exists a singleton set \(\{ A \}\) that contains only \(A\).

(e) For any nonempty family \(\mathcal{A}\), the intersection \(\bigcap \mathcal{A}\) exists.
(f) For any two sets $A$ and $B$, their intersection $A \cap B$ exists.

(g) For any family $\mathcal{A}$, the union $\bigcup \mathcal{A}$ exists.

(h) For any two sets $A$ and $B$, their union $A \cup B$ exists.

(i) For any two sets $A$ and $B$, their difference $A \setminus B$ exists.

(j) For any set $A$, its power set $\text{pow}(A)$ exists.

As a consequence, even subsets of $A$ which are not definable exist.

(k) For any set $A$, its successor $\text{succ}(A)$ exists.

(l) For any two sets $A$ and $B$, their Kuratowski pair $(A, B)$ exists.

(m) For any sets $\mathcal{K}$ and $A$, any indexed family $\{A_k\}_{k \in \mathcal{K}} \subseteq A$ exists.

(n) For any indexed family $\{A_k\}_{k \in \mathcal{K}}$, its Cartesian product $\prod_{k \in \mathcal{K}} A_k$ exists.

(o) For any two sets $A$ and $B$, the set of all relations between $A$ and $B$ exists.

(p) For any two sets $A$ and $B$, the set $\text{fun}(A, B)$ exists.

(q) For any set $A$ and any equivalence relation $\equiv$, the quotient set $A/\equiv$ exists.

(r) Fix some sets $A$ and $B$ and some indexed family of functions $\{f_k\}_{k \in \mathcal{K}}$ where $f : A \to B_k$ for $k \in \mathcal{K}$. For any $x \in A$ the corresponding tuple $\{f_k(x)\}_{k \in \mathcal{K}}$ exists.

Proof.

Proof of 985 (a). This is a trivial consequence of the axiom schema of specification.

Proof of 985 (b). As a consequence of the axiom of infinity, there exists at least one inductive set. Let $A$ be an inductive set. Then from the axiom schema of specification it follows that

$$\{x \in A \mid \bot\}$$

is a set. Furthermore, $x$ belongs to this set if and only if $x$ satisfies $\bot$, which is impossible, hence the set is empty.

As a consequence of the axiom of extensionality, this empty set is unique. As discussed in definition 931, we denote this unique empty set by $\emptyset$.

Proof of 985 (c). Aiming at a contradiction, suppose that there was a universal set $U$. Then we can easily reproduce theorem 932 (Russell’s paradox) by using restricted (to $U$) rather than unrestricted comprehension.

Unlike in naïve set theory, however, the existence of $U$ is not an axiom of the theory. Therefore, rather than demonstrating that ZFC is inconsistent, Russell’s paradox shows that certain sets like the universal set do not exist in ZFC.

Proof of 985 (d). Fix a set $A$. The set $\{A\}$, if it exists, is equal to $\{A\} = \{A, A\}$, by the axiom of extensionality.

Thus, by the axiom of pairing, the singleton set $\{A\} = \{A, A\}$ actually exists.
Proof of 985 (e). Let $\mathcal{A}$ be a nonempty family of sets. Their intersection $\bigcap \mathcal{A}$, if it exists, is a subset of every set $A \in \mathcal{A}$.

Therefore, since the family $\mathcal{A}$ is nonempty, the axiom schema of specification applied to any set in $A \in \mathcal{A}$ guarantees the existence of the intersection $\bigcap \mathcal{A}$. More precisely, for any $A_0 \in \mathcal{A}$, we can define the intersection of $A$ as

$$\bigcap \mathcal{A} = \{ x \in A_0 \mid \exists A \in \mathcal{A}. x \in A \}.$$ 

Proof of 985 (f). For sets $A$ and $B$, by the axiom of pairing the set $\{A, B\}$ exists. Then by proposition 985 (e), the binary intersection

$$A \cap B = \bigcap \{A, B\}$$

also exists.

Proof of 985 (g). The existence of arbitrary unions is merely a restatement of the axiom of unions.

Proof of 985 (h). Similarly to proposition 985 (f), for sets $A$ and $B$, by the axiom of pairing the set $\{A, B\}$ exists and by the axiom of unions, the binary union

$$A \cup B = \bigcup \{A, B\}$$

exists.

Proof of 985 (i). The difference $A \setminus B$ is guaranteed to exist by restricted comprehension:

$$A \setminus B = \{ x \in A \mid x \notin B \}.$$ 

Proof of 985 (j). The existence of power sets is a restatement of the axiom of power sets.

Proof of 985 (k). The successor of $A$ is

$$\text{succ}(A) = \{A\} \cup A.$$ 

Its existence follows from proposition 985 (d) and proposition 985 (h).

Proof of 985 (l). The existence of the Kuratowski pair

$$\langle A, B \rangle = \{\{A\}, \{A, B\}\}$$

can be proven by the axiom of pairing applied first to $\{A, B\}$ and then to the pair itself.

Proof of 985 (m). Given an indexed family $\{A_k\}_{k \in \mathcal{K}}$, by the axiom schema of replacement, there exists a set

$$\mathcal{A} := \{ A_k \mid k \in \mathcal{K} \}.$$ 

The family $\{A_k\}_{k \in \mathcal{K}}$ is, formally, a set of Kuratowski pairs. Every pair $(k, A_k)$ is itself a subset of $\text{pow}(\mathcal{K} \cup \mathcal{A})$. The family is then a subset of $\text{pow}(\text{pow}(\mathcal{K} \cup \mathcal{A}))$. Applying the axiom of power sets again, we obtain that the family exits as a set.
Proof of 985 (n). By definition, the Cartesian product \( \prod_{k \in \mathcal{K}} A_k \) is a set of indexed by \( \mathcal{K} \) families of members of the union \( \bigcup \{ A_k \mid k \in \mathcal{K} \} \).

We can apply the axiom of power sets one more time in addition to those in proposition 985 (m) to obtain the set of all indexed by \( \mathcal{K} \) families of members of this union. Then we can apply the axiom schema of specification to restrict only to those families that satisfy the condition of definition 950 (d).

Proof of 985 (o). All relations between \( A \) and \( B \) are subsets of \( A \times B \), hence elements of \( \text{pow}(A \times B) \). The latter exists by proposition 985 (n) and proposition 985 (j).

Proof of 985 (p). The set of single-valued functions from \( A \) to \( B \) is a subset of \( \text{pow}(A \times B) \), hence it exists by proposition 985 (o) and proposition 985 (j).

Proof of 985 (q). Let \( A \) be an arbitrary set and \( \equiv \) be a binary relation over \( A \). Then \( A/\equiv \) is a subset of \( \text{pow}(A) \) and hence it exists as a consequence of the axiom of power sets and the axiom schema of specification — see proposition 960 (b).

Proof of 985 (r). The set \( \{ f_k \}_{k \in \mathcal{K}} \) exists because it is a member of \( \text{fun}(\mathcal{K}, \text{fun}(A, B)) \), which set exists by proposition 985 (p).

Then \( \{ f_k(x) \}_{k \in \mathcal{K}} \) is the function

\[
\begin{align*}
g_x & : \text{fun}(\mathcal{K}, B) \\
g_x(k) & := f_k(x).
\end{align*}
\]

\( \square \)

Theorem 986 (Multi-valued selection existence). Every total multi-valued function has a selection.

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (d).

Proof.

Proof that the axiom of choice implies selection existence. Let \( F : A \rightrightarrows B \) be a total multi-valued function. As described in remark 975, we can instead take the indexed family \( \{ F(a) \}_{a \in A} \). Denote by \( f \) the single-valued function from \( A \) to the image \( \{ F(a) \mid a \in A \} \subseteq \text{pow}(B) \) of this indexed family (see remark 975 for clarifications).

Since \( F \) is total, the family \( \text{img}(f) = \{ F(a) \}_{a \in A} \) is a (potentially empty) family of nonempty sets. Thus, we can apply the axiom of choice to obtain a choice function \( c : \text{img}(f) \to B \).

The composition \( c \circ f \) is then a single-valued function. Furthermore, we have

\[
(c \circ f)(a) \in f(a) = F(a)
\]

so \( c \circ f \) is a selection of \( F \).

Proof that selection existence implies axiom of choice. Fix a family \( \mathcal{A} \) of nonempty sets. Define the function

\[
\begin{align*}
F : \mathcal{A} & \to \bigcup \mathcal{A} \\
F(A) & := A
\end{align*}
\]

502
that sends each set in $\mathcal{A}$ to the corresponding subset of $\mathcal{A}$. In terms of relations, we have $(A, x) \in F$ if and only if $x \in A$. This is a total multi-valued function because every set in $\mathcal{A}$ is nonempty.

Then every selection of $F$ is a choice function for $\mathcal{A}$.

**Definition 987.** The **disjoint union** of the indexed family $\{A_k\}_{k \in \mathcal{K}}$ of nonempty sets is

$$\bigsqcup_{k \in \mathcal{K}} A_k := \{(k, x) \mid k \in \mathcal{K} \text{ and } x \in A_k\}.$$  

**Theorem 988** (Surjective functions are right-invertible). Every surjective function is right-invertible.

Within $ZF$, this theorem is equivalent to the axiom of choice — see theorem 990 (e).

**Proof.** Let $f : A \to B$ be any function. Its inverse $f^{-1} : B \rightrightarrows A$ is, by definition, a partial multi-valued function.

**Proof that the axiom of choice implies right-invertibility.** If $f$ is surjective, then by definition 971 (b iii), its inverse is total. Then the axiom of choice via theorem 986 (Multi-valued selection existence) gives us a single-valued selection $g$ of $f^{-1}$.

Since the value of $f$ is $y$ for all members of $f^{-1}(y)$, and since $g(y) \in f^{-1}(y)$, we have

$$[f \circ g](y) = f(g(y)) = y.$$  

Therefore, $g$ is a right inverse of $f$.

**Proof that right-invertibility implies the axiom of choice.** Suppose that every surjective function is invertible.

Let $\mathcal{A}$ be an arbitrary family of nonempty sets. We can regard it as the indexed family $\{A\}_{A \in \mathcal{A}}$. Define the function

$$f : \bigsqcup_{A \in \mathcal{A}} A \to \mathcal{A},$$  

$$f(A, x) := A,$$  

where by $\bigsqcup$ we have denoted the disjoint union.

This function is surjective by definition. Then there exists a right-inverse

$$g : \mathcal{A} \to \{A\}_{A \in \mathcal{A}}.$$  

For every set $A \in \mathcal{A}$ we have $[f \circ g](A) = A$. Given a set $A$, $g$ gives us a pair $(A, x)$ with $x \in A$ and so $f(A, x) = A$.

Finally, define the choice function

$$c : \mathcal{A} \to \bigcup\mathcal{A},$$  

$$c(A) := x \text{ where } (A, x) = g(A).$$  

Since the family $\mathcal{A}$ was arbitrary, we can conclude that the axiom of choice holds. □
Proposition 989. We prove this result here rather than in section 13.3 (Functions) because it requires theorem 988 (Surjective functions are right-invertible), which requires the axiom of choice introduced in this section.

In relation to morphism invertibility in the category Set, we have the following:

(a) An empty function is always injective and left-invertible, however only if its range is empty is it left-cancellative, right-cancellative, right-invertible or surjective.

(b) A nonempty function is left-invertible if and only if it is injective.

(c) A function is left-cancellative if and only if it is injective.

(d) A function is right-invertible if and only if it is surjective.

(e) A function is right-cancellative if and only if it is surjective.

(f) A function is bijective if and only if it is fully invertible.

Proof.

Proof of 989 (a). Let \( g : \emptyset \to C \) be the empty function to \( C \). It is vacuously injective. It is also left-invertible because the only function that can be composed with \( g : \emptyset \to C \) from the left is the unique function from \( \emptyset \) to \( \emptyset \).

Clearly \( g : \emptyset \to C \) it is surjective if and only if \( C = \emptyset \).

For left-invertibility, note that \( g : \emptyset \to C \) composed with the function \( h : C \to D \) on the left is another empty function \( h \circ g : \emptyset \to D \). The latter is the identity \( \text{id}_\emptyset \) if and only if \( C = \emptyset \).

For right-invertibility, note that \( g : \emptyset \to C \) composed with the function \( f : A \to \emptyset \) on the right is the function \( g \circ f : A \to C \). But \( A = \emptyset \) since otherwise \( f \) would be nonempty, hence \( g \circ f : \emptyset \to C \). The latter is the identity \( \text{id}_\emptyset \) if and only if \( C = \emptyset \).

For right-cancellation, note that \( g \circ f_1 = g \circ f_2 \) implies \( f_1 = f_2 \) if and only if \( B \) is empty.

Proof of 989 (b). Let \( f : A \to B \) be a nonempty injective function. Definition 971 (a iii) states that the inverse \( f^{-1} : B \to A \) is a partial single-valued function.

Fix some value \( a \in A \) and define

\[
g : B \to A
\]

\[
g(y) := \begin{cases} f^{-1}(y), & y \in f(A) \\ a, & \text{otherwise.} \end{cases}
\]

This function \( g \) is a left inverse of \( f \) because, for any \( x \in A \),

\[
[g \circ f](x) = g(f(x)) = f^{-1}(f(x)) = x.
\]

We can see that \( g \) would be unique except for our choice of \( a \). We may even define \( g \) to take different values in \( A \) outside \( f(A) \). Thus, \( g \) is non-unique in general.

Conversely, suppose that \( f : A \to B \) is not necessarily injective and let \( g : B \to A \) be a left inverse of \( f \). Let \( x_1 \) and \( x_2 \) be two different points in \( A \). Since \( g \circ f = \text{id}_A \), clearly \( g(f(x_1)) \neq g(f(x_2)) \). If we suppose that \( f(x_1) = f(x_2) \), we would obtain a contradiction since then \( g(f(x_1)) \) would equal \( g(f(x_2)) \). Hence, \( f(x_1) \neq f(x_2) \). This shows that \( f \) is injective.
**Proof of 989 (c).** The case with an empty function is handled in proposition 989 (a), and we assume that it is nonempty.

Suppose that \( g : B \to C \) is a nonempty left-cancellative function. Let \( y_1 \) and \( y_2 \) be some members of \( B \) such that \( g(y_1) = g(y_2) \).

Suppose that \( y_1 \neq y_2 \) and define the function

\[
   f : B \to B
   f(y) := \begin{cases} 
   y_2, & y = y_1 \\
   y_1, & y = y_2 \\
   y, & y \neq y_1 \text{ and } y \neq y_2
   \end{cases}
\]

Then

\[
   g(f(y_2)) = g(y_1) = g(y_2) = g(f(y_1))
\]

For all \( y \in B \) different from \( y_1 \) and \( y_2 \), we have \( y = f(y) \).

Since \( g \) is left-cancellative, from \( g \circ \text{id}_B = g \circ f \) it follows that \( \text{id}_B = f \), which is a contradiction.

It remains for \( y_1 \) to be equal to \( y_2 \). Since these were arbitrary points in \( B \) satisfying \( g(y_1) = g(y_2) \), we conclude that \( g \) is injective.

Conversely, if \( f \) is injective, it is left-invertible by proposition 989 (b) and left-cancellative by proposition 1128 (a).

**Proof of 989 (d).** In one direction, we have theorem 988 (Surjective functions are right-invertible).

Conversely, suppose that \( g : B \to A \) is a right inverse of \( f : A \to B \). Let \( y \in B \). We have that \( g(y) \) is in the preimage of \( y \) under \( f \) because \( f(g(y)) = y \). Thus, the preimage is not empty for an arbitrary point in \( B \). We conclude that \( f \) is surjective.

**Proof of 989 (e).** The case with an empty function is handled in proposition 989 (a), and we assume that it is nonempty.

Let \( f : A \to B \) be a nonempty right-cancellative function. Suppose that it is not surjective and let \( y_0 \in B \setminus \text{img} f \). Let \( z \) be some set not belonging to \( B \). Define the function

\[
   g : B \to B \cup \{z\}
   g(y) := \begin{cases} 
   z, & y = y_0 \\
   y, & y \neq y_0
   \end{cases}
\]

Since \( f \) is right-cancellative, from \( \text{id}_B \circ f = g \circ f \) it follows that \( \text{id}_B = g \), which is a contradiction. Therefore, \( f \) is surjective.

We can prove the converse using proposition 989 (d) like we did for proposition 989 (c), however we prefer a direct proof that does not rely on the axiom of choice. As a bonus, this would allow us to prove theorem 1129 (Epimorphisms split in Set).

Conversely, suppose that \( f : A \to B \) is surjective and that for some functions \( g_1, g_2 : B \to C \) we have

\[
   g_1 \circ f = g_2 \circ f.
\]
Fix some \( y \in B \). Because \( f \) is surjective, there exists some \( x \in A \) such that \( f(x) = y \). Then

\[
g_1(y) = (g_1 \circ f)(x) = (g_2 \circ f)(x) = g_2(y).
\]

Since \( y \in B \) was arbitrary, we conclude that \( f \) is right-cancellative.

It is right-invertible by theorem 988 (Surjective functions are right-invertible) and right-cancellative by proposition 1128 (b).

**Proof of 989 (f).** If \( f : \emptyset \to B \) is a bijective empty function, then it is surjective and, by proposition 989 (d), it is right-invertible. By proposition 989 (a), it is also left-invertible. Thus, it is fully invertible.

Conversely, if an empty function \( f : \emptyset \to B \) is fully invertible, by proposition 989 (a) we have \( A = B = \emptyset \) and hence it is bijective.

Finally, if \( f : A \to B \) is nonempty function, then by proposition 989 (b) and proposition 989 (d) it is bijective if and only if it is fully invertible.

In the bijective case, we can avoid the axiom of choice via theorem 988 (Surjective functions are right-invertible) by noting that if \( f \) is bijective, its inverse is single-valued, and thus it is not necessary to do a selection of \( f^{-1} \).

**Theorem 990** (Axiom of choice equivalences). The following statements are commonly referred to as “the” axiom of choice:

(a) For every family of nonempty sets \( \mathcal{A} \) there exists a set \( B \) such that \( A \cap B \) is a singleton set for every \( A \in \mathcal{A} \).

(b) Every family of nonempty sets has a corresponding choice function.

(c) The Cartesian product of a family of nonempty sets is nonempty.

The following statements are equivalent to the axiom of choice, but are not conflated with it:

(d) **Theorem 986** (Multi-valued selection existence): Every total multi-valued function has a selection.

(e) **Theorem 988** (Surjective functions are right-invertible): Every surjective function is right-invertible.

(f) **Theorem 1312** (Hypergraph minimal transversal existence): Every hypergraph has a minimal transversal.

(g) **Theorem 1129** (Epimorphisms split in Set): Every epimorphism in \( \textbf{Set} \) splits.

(h) **Theorem 1175** (Fully faithful and essentially surjective functor induces equivalence): Every fully faithful and essentially surjective on objects functor induces a category equivalence.

(i) **Theorem 1183** (Category skeleton existence): Every category has a skeleton.

(j) **Theorem 1032** (Zermelo’s well-ordering theorem): Every set can be well-ordered.
Theorem 1240 (Zorn’s lemma): If every chain in a partially ordered set has an upper bound, then the entire set has a maximal element.

Theorem 561 (Maximal ideal theorem): Every proper semiring ideal is contained in a maximal ideal.

Theorem 626 (Vector space basis existence): Every vector space has a basis.

Theorem 321 (Tychonoff’s product theorem): The topological product of compact spaces is compact.

Proof. The equivalence proofs can be found in the linked theorems since that is usually the most appropriate place to put them. □

Theorem 991 (Diaconescu-Goodman-Myhill theorem). In ZF, the axiom of choice entails the law of the excluded middle (AX LEM).
13.5. Well-ordered sets

This section may fit in section 15 (Order theory), however it deals with well-foundedness of set membership and with induction principles which are more fitting here.

**Definition 992.** Let \( \prec \) be a binary relation on a set \( A \) (not necessarily satisfying any axioms). An **infinitely descending sequence** is a sequence \( \{x_k\}_{k=1}^{\infty} \) such that \( x_{k+1} \prec x_k \) for all \( k \in 1, 2, \ldots \). That is,

\[
\cdots \prec x_3 \prec x_2 \prec x_1 .
\]

The relation \( \prec \) is called **well-founded** if there exists no infinitely descending sequence.

We cannot easily formulate the theory of well-founded relations as a first-order theory. This is why we only discuss well-relations sets in the context of ZFC.

**Proposition 993.** Set membership is well-founded in ZFC. More precisely, given a set \( A \), if we regard \( \in \) as a binary relation between members of \( A \), then we would obtain that \( \in \) is a well-founded relation.

This proposition generalizes corollary 994.

**Proof.** The empty set is vacuously well-founded, so suppose that \( A \) is nonempty. Suppose also that \( \in \) is not well-founded on \( A \). Then there exists an infinitely descending sequence \( \{x_k\}_{k=1}^{\infty} \subseteq A \) such that

\[
\cdots \in x_3 \in x_2 \in x_1 .
\]

Denote by \( B \) the set \( \{x_k \mid k = 1, 2, \ldots \} \). By the **axiom of foundation**, \( B \) contains a set \( C \) which is disjoint from \( B \).

Clearly \( C \) coincides with at least one of \( x_1, x_2, \ldots \). Let \( C = x_{k_0} \). Since \( x_{k_0} \cap B = \emptyset \), then \( x_{k_0+1} \) is either not a member of \( x_{k_0} \) or of \( B \). But it is a member of both by our assumption that the sequence is infinitely descending.

The obtained contradiction proves that \( \in \) is well-founded on \( A \).

**Corollary 994.** We list several statements, which are consequences of the axiom of foundation via proposition 993:

(a) No set is a member of itself.

(b) For any two sets \( A \) and \( B \), it is not possible for both \( A \in B \) and \( B \in A \) to hold.

(c) The **ordinal successor operation** is injective. That is, for every two set \( A \) and \( B \) from \( \text{succ}(A) = \text{succ}(B) \) it follows that \( A = B \).

**Proof.**

**Proof of 994 (a).** The sequence

\[
\cdots A \in A \in A
\]

is infinitely descending, hence \( A \) cannot be a member of itself.
Proof of 994 (b). The sequence
\[\cdots A \in B \in A \in B \in A\]
is infinitely descending, hence \(A \in B\) and \(B \in A\) cannot simultaneously hold.

Proof of 994 (c). Note that \(A \in \text{succ}(A) = \text{succ}(B)\), hence either \(A = B\) or \(A \in B\).

We can analogously conclude that either \(A = B\) or \(B \in A\).

But \(A \in B\) and \(B \in A\) cannot simultaneously hold due to corollary 994 (b).

Hence, it remains for \(A = B\) to hold.

Proposition 995. In a partially ordered set \((P, \leq)\), the strict relation \(<\) is well-founded if and only if every nonempty subset \(A \subseteq P\) has a minimal element.

Proof.

Proof of sufficiency. Suppose that \(<\) is well-founded and let \(A \subseteq P\). Suppose that \(A\) has no minimal element.

If \(A\) is finite, we cannot construct an infinitely descending sequence because of irreflexivity of \(<\). If \(A\) is infinite, then we can construct an infinitely descending sequence using well-founded recursion as follows:

- Let \(x_1\) be any element of \(A\).
- Given \(x_k\), let \(x_{k+1}\) be any element of \(A\) such that \(x_{k+1} < x_k\). We know that such an element exists because \(x_k\) is not minimal in \(A\).

This construction contradicts the well-foundedness of \(\leq\), hence \(A\) must have a minimal element.

Proof of necessity. Suppose that every subset of \(P\) has a minimal element. Suppose that there exists an infinitely descending sequence \(\{x_k\}_{k=1}^{\infty}\). Then the set \(\{x_k \mid x \in \{1, 2, \ldots\}\}\) has no minimal element, which contradicts our assumption. Hence, no sequence in \(P\) is infinitely descending and \(<\) is well-founded.

Definition 996. A totally ordered set \((P, \leq)\) is said to be well-ordered if either of the following equivalent conditions hold:

(a) Every nonempty subset of \(P\) has a minimum.

(b) The strict order \(<\) is well-founded.

The irreflexivity of \(<\) is redundant because it follows from the well-foundedness — if the strict relation is not irreflexive, then there exists some element \(x \in P\) such that \(x < x\) and thus \(\{x\}_{k=1}^{\infty}\) is an infinitely descending sequence.

Proof. The equivalence of the conditions follows from proposition 995 and proposition 1245.
Theorem 997 (Well-founded induction). Let $\mathcal{L}$ be the first-order language with no functional symbols and a single predicate symbol $\prec$. We have already mentioned that we cannot formalize the concept of well-foundedness in first-order logic alone, so we will work with structures directly.

Every well-founded structure $X = (A, I)$ over $\mathcal{L}$ satisfies an inductive axiom schema. For every formula $\varphi$ over $\mathcal{L}$ not containing $\eta$ nor $\zeta$ as free variables, $X$ satisfies

$$\forall \eta. \left( (\forall \xi \prec \eta. \varphi[\xi \mapsto \zeta]) \rightarrow \varphi[\xi \mapsto \eta] \right) \rightarrow \forall \eta. \varphi[\xi \mapsto \eta]. \quad (353)$$

See the comments in definition 1 regarding variables and quantification in axiom schemas and remark 1027 for a general discussion of induction.

In the special case where $X = \mathbb{N}$, this is called strong induction compared to the usual natural number induction (PA3). This is discussed in remark 1027 (b).

Proof. Fix a formula $\varphi$ in $\mathcal{L}$. Fix a variable assignment $v$ in $X$. We will show that the contrapositive of (353) holds in this model.

Suppose that there exists a value $x \in A$ such that $\varphi[v_{\xi \mapsto x}] = F$. That is,

$$\exists \eta. \neg \varphi[\xi \mapsto \eta].$$

holds.

Let $x_0$ be the smallest such value (which is guaranteed to exist because $A$ is well-founded by $\prec$). Thus, the inductive hypothesis $\forall \xi < x_0. \varphi[\xi \mapsto \zeta]$ holds, but the inductive step conclusion $\varphi[\xi \mapsto x_0]$ does not.

Since $v$ was chosen arbitrarily, this is true for all variable assignments in $X$. Formally,

$$X \models \exists \eta. \neg \varphi[\xi \mapsto \eta] \rightarrow \exists \eta. \left( \forall \xi < \eta. \varphi[\xi \mapsto \zeta] \wedge \neg \varphi[\xi \mapsto \eta] \right) \quad (354)$$

This is precisely the contrapositive of (353). Since we are working in classical logic, the contrapositive is semantically equivalent to its original implication, hence $X \models (353)$. $\square$

Theorem 998 (Epsilon induction). For every formula $\varphi$ in the language of set theory not containing $\eta$ nor $\zeta$ as free variables, the following is a theorem of ZFC:

$$\forall \eta. \left( (\forall \xi \in \eta. \varphi[\xi \mapsto \zeta]) \rightarrow \varphi[\xi \mapsto \eta] \right) \rightarrow \forall \eta. \varphi[\xi \mapsto \eta].$$

This induction schema is called “$\varepsilon$-induction” because the set membership symbol $\in$ is derived from $\varepsilon$ as explained in remark 925.

See the comments in definition 1 regarding variables and quantification in axiom schemas and remark 1027 for a general discussion of induction.
Proof. Every model of ZFC is well-founded by ∈ due to proposition 993. The corollary then follows from theorem 997 (Well-founded induction).

Lemma 999. Any order embedding on a well-ordered set is inflationary.
That is, if \((P, \leq)\) is a well-ordered set and \(f : P \to P\) is an order embedding, then \(x \leq f(x)\) for any \(x \in P\).

Proof. We proceed by induction on \(P\). Fix \(x_0\) and suppose that \(y \leq f(y)\) for all \(y < x_0\).

Aiming at a contradiction, suppose that \(f(x_0) < x_0\). Thus, there exists some \(y_0 < x_0\) such that \(f(x_0) = y_0\). By the inductive hypothesis we have \(f(x_0) = y_0 \leq f(y_0)\). Thus, either \(f(x_0) < f(y_0)\), which contradicts that \(f\) is an order homomorphism, or \(f(x_0) = f(y_0)\), which contradicts the injectivity of \(f\).

The obtained contradiction demonstrates that \(x \leq f(x)\) for all \(x \in P\).

Proposition 1000. There is at most one isomorphism between any pair of well-ordered sets.

Proof. Fix two well-ordered sets \((P, \leq_P)\) and \((Q, \leq_Q)\) and let \(f : P \to Q\) and \(g : P \to Q\) be two order isomorphisms.

For any member \(x\) of \(P\) we have the following possibilities:

- If \(f(x) <_Q g(x)\), then \(g^{-1}(f(x)) <_P x\) and \(f(g^{-1}(f(x))) <_Q f(x)\). Then we can use natural number recursion to build an infinitely descending sequence of members of \(Q\).
  
  But this is a contradiction because \(Q\) is well-founded.

- We can build a similar sequence if \(g(x) <_Q f(x)\).

- It remains for \(f(x) = g(x)\) to hold.

Since \(x \in P\) was arbitrary, we conclude that \(f = g\).

Proposition 1001. If \((P, \leq_P)\) and \((Q, \leq_Q)\) are well-ordered sets, then the lexicographic and reverse lexicographic orders on \(P \times Q\) are well-ordering relations.

Compare this result to proposition 1236 and proposition 1248.

Proof. We have already shown in proposition 1248 and these total orders. It only remains to check well-foundedness.

Proof of well-foundedness. Let \(<\) be the lexicographic order on \(P \times Q\).

Suppose that there exists an infinitely descending sequence

\[
\cdots < (a_3, b_3) < (a_2, b_2) < (a_1, b_1).
\]

Since \(P\) is well-founded, the corresponding sequence \(\{a_k\}_{k=1}^\infty\) s.t. \(a_{k_0} = a_k\) for \(k \geq k_0\). Therefore, the sequence \(\{b_k\}_{k=k_0}^\infty\) must be infinitely descending. But this contradicts the well-foundedness of \(Q\).

Therefore, no infinitely descending sequence exists in the totally ordered set \((P \times Q, <)\) and thus it is well-ordered.
13.6. Ordinals

Remark 1002. Ordinals are generalizations of natural numbers. We will find characterizing properties of the natural numbers (defined as members of $\omega$), so that it is clear what we want to generalize.

Every natural number is defined as a set of other natural numbers:

\[
\begin{align*}
0 &= \emptyset \\
1 &= \{\emptyset\} = \{0\} \\
2 &= \{\emptyset, \{\emptyset\}\} = \{0, 1\} \\
3 &= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0, 1, 2\}
\end{align*}
\]

It just so happens that each natural number $n$ is the set of natural numbers that are smaller with respect to the strict order relation $<$ defined in (3).

Therefore, $\in$ and $<$ are equivalent on the set $\omega$. It follows from proposition 11 that $\in$ is a strict total order on $\omega$. It is even well-ordered by $\in$, but the latter condition is redundant due to proposition 993.

For an arbitrary set $A$, set membership is not even a strict partial order — irreflexivity is implied by proposition 993, but transitivity of $\in$ as a binary relation on $A$ fails to hold in general, not speaking about trichotomy.

A very simple counterexample for transitivity of $\in$ is the set $\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}\}$. Clearly $\emptyset \in \{\emptyset\}$ and $\{\emptyset\} \in \{\{\emptyset\}\}$, but $\emptyset \not\in \{\{\emptyset\}\}$.

In order for a set $A$ to be a member of $\omega$, it is not sufficient for $\in$ to be a strict total ordered on $A$. Except for the members of $\omega$, another set that is totally ordered by $\in$ is $A = \{0, 2, 4\}$.

If we require $A = \{0, 2, 4\}$ to be a transitive set, however, it will be a natural number. Indeed, since 4 is a member of $A$ and 1 and 3 are members of 4, then by adding 1 and 3 to $A$ we obtain the set $\{0, 1, 2, 3, 4\}$, which by our definition of natural numbers is 5.

Note that transitivity of the relation $\in$ on $A$ and transitivity of the set $A$ itself are entirely different concepts, although we will use both. Every member of $\omega$ is a transitive set by proposition 945 and the relation $\in$ is a strict total order by proposition 11.

This is the reasoning behind our definition of an ordinal — definition 1003. From this definition it will follow that the ordinals are unique representatives of order-isomorphisms classes of well-ordered sets.

As a final note, the above two conditions are not sufficient for $A$ to be a member of $\omega$ (they are too general), but if we additionally require $A$ to be a finite set, then $A$ will be a member of $\omega$. We have yet to define finiteness, however.

Definition 1003. An ordinal number or simply ordinal is a transitive set $A$ such that set membership (as a binary relation on $A$) well-orders $A$. By tradition, ordinals are denoted by initial small Greek letters like $\alpha$ and $\beta$.

Because of proposition 993, it is sufficient for set membership to be a strict total order on $A$. Since well-foundedness also implies irreflexivity, it follows that set membership must only be transitive and trichotomic on $A$.

In the absence of the axiom of foundation, we additionally require set membership to be a well-founded relation on $A$, so that $A$ is well-ordered.
See remark 1002 for a further discussion of the definition, especially the different notions of transitivity.

We introduce the notation $\alpha < \beta$ for $\alpha \in \beta$ in analogy with natural numbers. This is not a binary relation since there is no set of all ordinals by theorem 1019 (Burali-Forti paradox), however it does satisfy the properties of a well-order due to proposition 1012 (b).

Finally, we introduce the following predicate formula

$$
\text{IsOrdinal}[\tau] := \text{IsSetTransitive}[\tau] \land
\left( \forall \xi \in \tau. \forall \eta \in \tau. \left( \eta \in \xi \lor \eta = \xi \lor \xi \in \eta \right) \land
\left( \forall \xi \in \tau. \forall \eta \in \tau. \forall \zeta \in \tau. \left( (\xi \in \eta \land \eta \in \zeta) \rightarrow \xi \in \zeta \right) \right) \right).
$$

**Proposition 1004.** The smallest inductive set $\omega$ is an ordinal.

**Proof.** From proposition 945 it follows that $\omega$ is a transitive set.

Also, as discussed in remark 1002, from proposition 11 it follows that set membership is a strict total order on $\omega$.

Therefore, $\omega$ is an ordinal. $\Box$

**Proposition 1005.** Every member of an ordinal is an ordinal.

**Proof.** Let $\alpha$ be an ordinal and let $\beta \in \alpha$. We will show that $\beta$ is an ordinal.

By transitivity of $\alpha$, we have $\beta \subseteq \alpha$ and by definition 1222 (e), $(\beta, \in)$ is a (strictly) totally ordered set as a substructure of $(\alpha, \in)$.

It remains to show that $\beta$ is itself transitive. Let $x \in \beta$. We have that $\beta \subseteq \alpha$ since $\alpha$ is transitive, hence $x \in \alpha$.

Fix $y \in x$. Again from the transitivity of $\alpha$ it follows that $y \in \alpha$. Also, $\in$ is a total order on $\alpha$ and hence from $y \in x$ and $x \in \beta$ it follows that $y \in \beta$.

Since $y \in x$ was chosen arbitrarily, it follows that $x \subseteq \beta$. Since $x$ was chosen arbitrarily, it follows that $\beta$ is transitive. $\Box$

**Proposition 1006.** Let $\alpha$ be an ordinal. For any $\beta \in \alpha$, the open initial segment $\alpha_{< \beta}$ equals $\beta$.

This is the bounded version of proposition 1018.

**Proof.** Let $\beta \in \alpha$. Consider the initial segment

$$
\alpha_{< \beta} = \{ y \in \alpha \mid y \in \beta \}.
$$

Clearly $\alpha_{< \beta} = \alpha \cap \beta$. Given that $\alpha$ is a transitive set, however, we have $\beta \subseteq \alpha$ and thus $\alpha \cap \beta = \beta$.

Therefore, $\alpha_{< \beta} = \beta$. $\Box$

**Corollary 1007.** The natural numbers (as members of $\omega$) are ordinals.

**Proof.** Follows from proposition 1004 and proposition 1005. $\Box$
Definition 1008. For any ordinal \(\alpha\) we call any function with \(\alpha\) as its domain a \(\alpha\)-indexed transfinite sequence.

In particular, the case \(\alpha = \omega\) corresponds to the standard natural number sequences.

Theorem 1009 (Bounded transfinite induction). For every formula \(\varphi\) in the language of set theory not containing \(\tau, \eta\) nor \(\zeta\) as free variables, the following is a theorem of ZFC:

\[
\forall \tau. \left( \text{IsOrdinal}[\tau] \rightarrow \left( \forall \eta \in \tau. \left( \forall \zeta \in \eta. \varphi[\xi \mapsto \zeta] \right) \rightarrow \varphi[\xi \mapsto \eta] \right) \rightarrow \forall \eta \in \tau. \varphi[\xi \mapsto \eta] \right).
\]

See the comments in definition 1 regarding variables and quantification in axiom schemas and remark 1027 for a general discussion of induction.

See remark 1024 about a reformulation that is often useful in practice.

Proof. This theorem is a special case of theorem 998 (Epsilon induction) with the formula \(\xi \in \tau \rightarrow \varphi\) that is explicitly universally quantified by the parameter \(\tau\) which ranges over all ordinals.

Note that it is unnecessary to verify that \(\eta\) and \(\zeta\) are ordinals because proposition 1005 ensures that \(\eta\) is only quantified over ordinals. \(\square\)

Theorem 1010 (Bounded transfinite recursion). Fix an ordinal \(\alpha\) and a nonempty set \(A\).

Suppose that we are given some transformation \(T : \text{pow}(\alpha \times A) \rightarrow A\) whose value for any relation between \(\alpha\). Then there exists a unique \(\alpha\)-indexed transfinite sequence \(f : \alpha \rightarrow A\) such that for any \(\beta \in \alpha\) we have \(f(\beta) = T(f|_{\beta})\).

This is a vast generalization of theorem 979 (Recursion theorem) from natural number sequences to transfinite sequences.

See remark 1024 about a reformulation that is often useful in practice.

Proof. The proof is analogous to that of theorem 979 (Recursion theorem), but we will give it anyway to highlight the difference between using theorem 943 (Induction via inductive sets) and theorem 1009 (Bounded transfinite induction).

Let \(G \subseteq \text{pow}(\alpha \times A)\) be the set of all partial single-valued functions \(g : \alpha \rightarrow A\) such that

- There exists some \(\beta_g \in \alpha\) such that \(g\) is defined only in the open initial segment \(\alpha_{<\beta_g}\). That is, \(g\) is defined for all \(\beta\) up to not including \(\beta_g\).
- \(g(\beta) = T(g|_{\beta})\) for all \(\beta < \beta_g\).

Clearly \(G\) is nonempty because the function \(\{ (\emptyset, T(\emptyset)) \}\) belongs to \(G\).

Define \(f := \bigcup G\). At this point \(f\) is a multi-valued function. We must now show that \(f\) has all the properties that we want.

Proof of totality. First, we will use theorem 1009 (Bounded transfinite induction) to show that \(f\) is total.

Fix \(\beta \in \text{dom } f\). Then there exists a function \(g \in G\) defined for all \(\gamma < \beta\).
If $g$ is also defined at $\beta$ also, this directly proves that $\beta \in \text{dom} \ f$.

If $g$ is not defined at $\beta$, consider

$$\hat{g} := g \cup \{(\beta, T(g|_{\beta}))\}.$$ 

The function $\hat{g}$ is again a single-valued partial function and thus it belongs to $G$, hence $\beta \in \text{dom} \ f$.

Therefore, theorem 1009 (Bounded transfinite induction) allows us to conclude that $f : \alpha \rightarrow A$ is a total multi-valued function.

**Proof of single-valuedness.** Now that we know that $f$ is total, we will prove that it is single-valued and thus is a function in the usual sense of the term.

Fix $\beta \in \alpha$. Suppose that $f$ is single-valued for all $\gamma < \beta$. Since $f$ is total, there exist at least one partial function $g$ in $G$ that is defined at $\beta$. Let $g$ and $h$ both be such (single-valued partial) functions.

Then

$$g(\beta) = T(g|_{\beta}) = T(f|_{\beta}) = T(h|_{\beta}) = h(\beta),$$

hence $g$ and $h$ coincide at $\beta$, which in turn implies that $f$ is single-valued at $\beta$.

Therefore, theorem 943 (Induction via inductive sets) allows us to conclude that $f$ is a single-valued total function.

**Proof of uniqueness.** Now that it is clear that $f$ satisfies the theorem, we must verify that it is unique.

Suppose that $f_1$ and $f_2$ both satisfy the theorem. Fix some $\beta \in \alpha$ and suppose that for every $\gamma < \beta$, we have $f_1(\gamma) = f_2(\gamma)$. Then

$$f_2(\beta) = T(f_1|_{\beta}) = T(f_2|_{\beta}) = f_2(\beta).$$

Therefore, theorem 943 (Induction via inductive sets) allows us to conclude that $f_1 = f_2$. So there is at most one function that satisfies the theorem and we have already shown that $f$ is such a function.

---

**Theorem 1011 (Transfinite induction).** It turns out that theorem 1010 (Bounded transfinite recursion) is valid for all ordinals simultaneously.

For every formula $\varphi$ in the language of set theory not containing $\eta$ nor $\zeta$ as free variables, the following is a theorem of ZFC:

\[
\forall \eta. \left( \text{IsOrdinal}[\eta] \rightarrow \left( \forall \zeta \in \eta. \varphi[\zeta \rightarrow \zeta] \right) \rightarrow \varphi[\xi \rightarrow \eta] \right) \rightarrow \forall \eta. \left( \text{IsOrdinal}[\eta] \rightarrow \varphi[\xi \rightarrow \eta] \right).
\]

This theorem could be a special case of theorem 1010 (Bounded transfinite recursion), but there exists no set of all ordinals due to theorem 1019 (Burali-Forti paradox).

See the comments in definition 1 regarding variables and quantification in axiom schemas and remark 1027 for a general discussion of induction.

See remark 1024 about a reformulation that is often useful in practice.
Proof. The proof is similar to the proof of theorem 997 (Well-founded induction).

Proposition 1012. The ordinals are well-ordered. Since there exists no set of all ordinals due to theorem 1019 (Burali-Forti paradox), we cannot say that the ordinals form a well-ordered set. We will instead state a more concrete result.

(a) For any three ordinals \( \alpha, \beta \) and \( \gamma \) such that \( \alpha < \beta < \gamma \) we have \( \alpha < \gamma \).

(b) For any two ordinals \( \alpha \) and \( \beta \), exactly one of \( \alpha = \beta \), \( \alpha < \beta \) or \( \alpha > \beta \) holds.

As discussed in definition 1003, irreflexivity and well-foundedness hold by proposition 993.

Proof.

Proof of 1012 (a). Let \( \alpha, \beta \) and \( \gamma \) be ordinals and let \( \alpha \in \beta \in \gamma \). Since \( \gamma \) is a transitive set, it follows that \( \beta \subseteq \gamma \) and thus \( \alpha \in \gamma \).

Therefore, we have used the fact that \( \gamma \) is a transitive set to prove that set membership is a transitive relation, thus obtaining a connection between two distinct concepts both named “transitivity”.

Proof of 1012 (b). Let \( \alpha \) and \( \beta \) be ordinals.

Due to corollary 994 (a), it is not possible for both \( \alpha \in \beta \) and \( \alpha = \beta \) to hold simultaneously. Due to corollary 994 (b), it is not possible for both \( \alpha \in \beta \) and \( \beta \in \alpha \) to hold simultaneously.

Thus, at most one of \( \alpha = \beta \), \( \alpha \in \beta \) or \( \beta \in \alpha \) holds.

We will use theorem 1011 (Transfinite induction) on \( \beta \) to show that at least one of them holds. Fix an ordinal \( \beta_0 \). Our inductive hypothesis is that for every ordinal \( \alpha \) and every \( \gamma \in \beta_0 \) at least one of \( \alpha = \gamma \), \( \alpha \in \gamma \) or \( \gamma \in \alpha \) holds.

Fix some ordinal \( \alpha_0 \). We will show that at least one of \( \alpha_0 = \beta_0 \), \( \alpha_0 \in \beta_0 \) or \( \beta_0 \in \alpha_0 \) holds. Since the case \( \beta_0 = \alpha_0 \) is trivial, we exclude it from consideration.

- If there exists \( \gamma \in \beta_0 \) such that \( \gamma = \alpha_0 \), clearly \( \alpha_0 \in \beta_0 \).
- If there exists \( \gamma \in \beta_0 \) such that \( \alpha_0 \in \gamma \), then by transitivity \( \alpha_0 \in \beta_0 \).
- If for every \( \gamma \in \beta_0 \) we have \( \gamma \in \alpha_0 \), then \( \beta_0 \subseteq \alpha_0 \). Let \( \gamma_0 \) be the smallest member of \( \alpha_0 \setminus \beta_0 \). We will show that \( \gamma_0 = \beta_0 \).

Our first goal is to show that \( \gamma_0 \subseteq \beta_0 \). Aiming at a contradiction, suppose that there exists some \( \lambda \in \gamma_0 \setminus \beta_0 \). Since \( \gamma_0 \) is a transitive set, we have \( \lambda \in \alpha \). Thus, \( \lambda \in \alpha \setminus \beta_0 \) and \( \lambda \in \gamma_0 \), contradicting the minimality of \( \gamma_0 \). Therefore, \( \gamma_0 \subseteq \beta_0 \).

Now we will use the existing inductive hypothesis for \( \alpha = \gamma_0 \) to show that \( \beta_0 \subseteq \gamma_0 \).

- If there exists \( \lambda \in \beta_0 \) such that \( \lambda = \gamma_0 \), clearly \( \gamma_0 \in \beta_0 \). But that contradicts our choice of \( \gamma_0 \) as a member of \( \alpha_0 \setminus \beta_0 \).
- If there exists \( \lambda \in \beta_0 \) such that \( \gamma_0 \in \lambda \), then by transitivity \( \gamma_0 \in \beta_0 \), which again contradicts our choice of \( \gamma_0 \).
- Finally, if for every \( \gamma \in \beta_0 \) we have \( \gamma \in \gamma_0 \), then \( \beta_0 \subseteq \gamma_0 \).
Thus, both \( \gamma_0 \subseteq \beta_0 \) and \( \beta_0 \subseteq \gamma_0 \), giving us the equality \( \beta_0 = \gamma_0 \). This allows us to conclude that \( \beta_0 \in \alpha_0 \).

We have shown by transfinite induction that for a fixed ordinal \( \beta_0 \), for every other ordinal \( \alpha \) at least one of \( \beta_0 = \alpha, \beta_0 \in \alpha \) or \( \alpha \in \beta_0 \) holds. We have already shown that at most one of the three holds, hence exactly one of the three holds.

Since \( \beta_0 \) is itself arbitrary, we can conclude that trichotomy holds for any two ordinals \( \alpha \) and \( \beta \).

**Proposition 1013.** For any two ordinals \( \alpha \) and \( \beta \) we have \( \beta \in \alpha \) if and only if \( \beta \subsetneq \alpha \).

**Proof.**

**Proof of necessity.** Since \( \alpha \) is a transitive set, from \( \beta \in \alpha \) it follows that \( \beta \subseteq \alpha \).

We cannot have \( \beta = \alpha \) due to corollary 994 (a), hence \( \beta \subsetneq \alpha \).

**Proof of sufficiency.** Suppose that \( \beta \subsetneq \alpha \).

By proposition 1012 (b), the ordinals \( \beta \) and \( \alpha \) must be related by set membership.

- If \( \alpha = \beta \), this directly contradicts our assumption that \( \beta \subsetneq \alpha \).
- If \( \alpha \in \beta \), then \( \alpha \in \alpha \) which contradicts corollary 994 (a).
- It remains for \( \beta \in \alpha \) to hold.

**Proposition 1014.** The ordinal successor operation is strictly monotone on ordinals. That is, if \( \alpha < \beta \), then \( \text{succ}(\alpha) < \text{succ}(\beta) \).

**Proof.** Let \( \alpha \in \beta \) and let \( \gamma \in \text{succ}(\alpha) \).

- If \( \gamma \in \alpha \), clearly \( \gamma \in \beta \) because \( \beta \) is a transitive set.
- If \( \gamma = \alpha \), then \( \gamma = \alpha \in \beta \).

We have shown that \( \text{succ}(\alpha) \subseteq \beta \). Thus, we either have \( \text{succ}(\alpha) = \beta \in \text{succ}(\beta) \) or else by proposition 1013 we have \( \text{succ}(\alpha) \in \beta \in \text{succ}(\beta) \).

**Proposition 1015.** A transitive set whose members are transitive sets is an ordinal.

In particular, a transitive set of ordinals is an ordinal. If a set of ordinals is not transitive, we can instead take its transitive closure.

**Proof.** Let \( A \) be a set whose members are themselves transitive sets.

We will first show that set membership is transitive on \( A \). If \( x, y \) and \( z \) are members of \( A \) such that \( x \in y \in z \), then since \( z \) is transitive we have \( y \subseteq z \) and thus \( x \in z \).

Therefore, we can conclude that set membership is a strict partial order on \( A \). Now define the set

\[
B := \{ x \in A \mid \exists y \in A . x \notin y \land y \notin x \}
\]
of all members of $A$ which are not related to at least one other member. If $B$ is empty, then set membership is trichotomous on $A$.

If $B$ is nonempty, let $b$ be a minimal element of $B$. A minimal element must exist by proposition 995 because set membership is a well-founded partial order on $A$. We have chosen $b$, so that every member of $b$ is related to every other member of $A$.

Define the set
$$C := \{x \in A \mid x \neq b \land x \not\in b \land b \not\in x\}$$
of all members of $A$ which are not related to $b$ and let $c$ be a minimal element of $C$. We will now show that $b = c$, which is a contradiction with our choice of $c$.

Let $x \in b$. As we have already mentioned, $x$ is related to every other member of $A$, including $c$.

- If $c = x$, then $c \in b$, which contradicts our choice of $c$.
- Suppose that $c \in x$. We have chosen $x$ to be a member of $b$ and we thus have $c \in x \in b$. The set $b$ is transitive as a member of $A$, hence $x$ is also a member of $A$. Since set membership is a transitive relation on $A$, it follows that $c \in b$, which contradicts our choice of $c$.
- It remains for $x \in c$ to hold.

Therefore, $b \subseteq c$. The converse inclusion $c \subseteq b$ can be obtained analogously by noting that $c$ is a minimal element of $C$ and hence every $x \in c$ is related to $b$. Thus, we obtain $b = c$, which contradicts our choice of $c$ as a member of $C$.

The obtained contradiction shows that the set $B$ is empty and thus every member of $A$ is related to every other member, proving trichotomy of set membership on $A$. This allows us to conclude that $A$ is an ordinal because it is a transitive set and we have already shown in the beginning of the proof that set membership is a transitive relation on $A$.

[Aut20] Definition 1016. To show that every set has a rank we must introduce additional definitions. We use natural number recursion to define the transitive closure of a set $A$ as
$$\text{cl}_T(A) := \bigcup_{n \in \omega} \text{cl}_n(A)$$

where
$$\text{cl}_n(A) := \begin{cases} A, & n = 0 \\ \bigcup \text{cl}_{n-1}(A), & n > 0 \end{cases}$$

Note that this is different from the transitive closure of a relation defined in definition 961 (c).

[Aut20] Proposition 1017. The transitive closure $\text{cl}_T(A)$ of the set $A$ is the smallest transitive set containing $A$.

Proof. It is clear that $A = \text{cl}_0^T(A)$ is a subset of $\text{cl}_T^T(A)$.

If $x \in \text{cl}_n^T(A)$, then there exists some natural number $n$ for which $x \in \text{cl}_n^T(A)$. Therefore, $x \subseteq \bigcup \text{cl}_n^T(A) = \text{cl}_{n+1}(A)$ and thus $x \subseteq \text{cl}_T^T(A)$.

Now suppose that $B$ is a transitive subset of $\text{cl}_T(A)$ which contains $A$. Let $x_0 \in \text{cl}_T(A)$. 518
Suppose that \( x_0 \not\in B \). Then there must exist a smallest nonzero number \( n \) such that \( x_0 \in \overline{T_n(A)} \). Then \( x_0 \) belongs to some member \( x_1 \) of \( \overline{T_{n-1}(A)} \). If \( x_1 \in A \), then \( x_0 \) must belong to \( B \) since it is transitive. But this contradicts our choice of \( x_0 \). Then \( x_1 \not\in A \), in which case there exists some \( x_2 \in \overline{T_{n-2}(A)} \) such that \( x_1 \in x_2 \). If \( x_2 \in A \), then again \( x_0 \not\in B \), which contradicts our choice of \( x_0 \). We can thus recursively construct a sequence \( \{x_k\}_{k=0}^{\infty} \) such that for every \( k \geq 0 \) both \( x_k \in x_{k+1} \) and \( x_k \not\in A \) hold. The existence of such a sequence contradicts Proposition 993.

Therefore, \( \overline{T(A)} \subseteq B \), hence \( \overline{T(A)} \) is the smallest transitive set containing \( A \).

**Proposition 1018.** Every ordinal equals the set of all smaller ordinals.

This is the unbounded version of Proposition 1006.

**Proof.** Let \( \alpha \) be an ordinal and let \( A \) be the set of all ordinals smaller than \( \alpha \). We will show that \( A = \alpha \). We will first show that \( A \) is a transitive set. Let \( \beta \in A \) and \( \gamma \in \beta \). Since \( \alpha \) is a transitive set that contains \( \beta \), we have \( \gamma \in \alpha \). Thus, \( \gamma \) is smaller than \( \alpha \) and hence it belongs to \( A \). Therefore, \( A \) is a transitive set of ordinals and by Proposition 1015, it is itself an ordinal. Proposition 1012 implies that \( \alpha \) and \( A \) are either equal or related by set membership.

- If \( \alpha \in A \), then \( \alpha \) is smaller than itself, which contradicts Corollary 994 (a).
- If \( A \in \alpha \), then \( A \) is smaller than itself, which again contradicts Corollary 994 (a).
- It remains for \( A \) to be equal to \( \alpha \).

**Theorem 1019** (Burali-Forti paradox). Assuming ZFC, there is no set of all ordinals.

**Proof.** Aiming at a contradiction, suppose that \( A \) is a containing all ordinals. If \( \alpha \in A \) and \( \beta \in \alpha \), transitivity \( \beta \) implies \( \beta \in A \). Thus, \( A \) is a transitive set of ordinals, which Proposition 1015 is itself an ordinal. Hence, \( A \in A \).

But this contradicts Corollary 994 (a). Hence, there is no set of all ordinals.

**Proposition 1020.** Let \( A \) be a set of ordinals and denote \( \alpha := \bigcup A \).

(a) The union \( \alpha \) is itself an ordinal.

(b) The union \( \alpha \), which is the supremum of \( A \) with respect to set inclusion, is also the supremum of \( A \) with respect to ordinal ordering.

That is, either \( \alpha = A \) or \( \alpha \) is the smallest ordinal that is larger than every member of \( A \).

(c) If \( A \) is an ordinal, then \( \alpha \leq A \). Furthermore, in the case \( \alpha < A \), there is no ordinal between \( A \) and \( \alpha \). That is, if \( \alpha < A \), then \( A \) is the smallest ordinal strictly larger than \( \alpha \).

See Definition 1023 (c) for a further distinction between \( \alpha = A \) and \( \alpha < A \).

For recursive definitions like Definition 1068 (a) this proposition justifies using \( \sup A \) instead of the more confusing \( \bigcup A \).

Compare this result to Proposition 1052.
Proof.

**Proof of 1020 (a).** Let $A$ be a set of ordinals. Denote its union by $\alpha := \bigcup A$.

We will show that $\alpha$ is a transitive set. Due to proposition 1015, this is sufficient for $\alpha$ to be an ordinal.

Let $\beta \in \alpha$. Then there exists an ordinal $\gamma$ in $A$ such that $\beta \subseteq \gamma$. Since $\gamma$ is itself a transitive set, we have $\beta \subseteq \gamma$. But $\gamma \subseteq \alpha$, hence $\beta \subseteq \alpha$.

Therefore, $\alpha = \bigcup A$ is a transitive set and thus an ordinal.

**Proof of 1020 (b).** For every $\beta \in A$ we have $\beta \subseteq \alpha$, which by proposition 1013 corresponds to $\beta \leq \alpha$ with respect to ordinal ordering.

We will show that $\alpha$ is the smallest ordinal with this property. Let $\lambda_0$ be any other ordinal such that $\gamma \leq \lambda_0$ for any $\gamma \in A$. Fix some $\gamma_0 \in \alpha$. Then there exists an ordinal $\beta_0 \in A$ such that $\gamma_0 \in \beta_0$. Since $\beta_0 \in A$ and $A \subseteq \lambda_0$, since $\lambda_0$ itself is a transitive set we have $\gamma_0 \in \lambda_0$.

Therefore, $\alpha$ is the least upper bound of $A$ with respect to ordinal ordering.

**Proof of 1020 (c).** Assume that $A$ is an ordinal.

If $\gamma \in \alpha$, by transitivity of $\alpha$ we have $\gamma \subseteq \alpha$. Thus, $\gamma \in A$. Hence, $\alpha \subseteq A$.

Now suppose that there exists some $\gamma \in A$ such that $\alpha \in \gamma$. Then

$$\gamma \subseteq \bigcup A = \alpha \in \gamma,$$

which contradicts corollary 994 (a).

Hence, there is not ordinal between $\alpha = \bigcup A$ and $A$. \qed

**Proposition 1021.** The successor $\alpha := \text{succ}(\beta)$ of an ordinal $\beta$ is the smallest ordinal larger than $\alpha$.

**Proof.** We must show that $\alpha$ is a transitive set and thus by proposition 1015 an ordinal. Note that $\alpha = \text{succ}(\beta) = \beta \cup \{\beta\}$.

Let $\gamma \in \alpha$.

- If $\gamma \in \beta$, then since $\beta$ is a transitive set, we have $\gamma \subseteq \beta$. Furthermore, since $\beta \subseteq \alpha$, by transitivity of set inclusion $\gamma \subseteq \alpha$.
- If $\gamma = \beta$, then $\gamma \subseteq \alpha$ by definition of successor.

Therefore, $\alpha$ is a transitive set and thus an ordinal.

Now suppose that there is another ordinal $\gamma$ such that $\beta \in \gamma \in \alpha = \beta \cup \{\beta\}$.

- If $\gamma \in \beta$, this would contradict corollary 994 (b).
- If $\gamma = \beta$, this would contradict corollary 994 (a).

The obtained contradictions show that there is no ordinal between $\alpha$ and $\beta$. \qed

**Remark 1022.** It follows from proposition 1018 and proposition 1021 and that for any ordinal $\beta$ we have

$$\beta = \{\gamma \mid \gamma \text{ is an ordinal and } \gamma < \beta\},$$

$$\text{succ}(\beta) = \{\gamma \mid \gamma \text{ is an ordinal and } \gamma \leq \beta\}.$$

This shows that the ordinal successor operation is very natural in the context of ordinals.
**Definition 1023.** We say that the ordinal $\alpha$ is a **successor ordinal** if any of the following equivalent conditions hold:

(a) The ordinal $\alpha$ is the successor of another ordinal. That is, there exists another ordinal $\beta$ such that $\alpha = \text{succ}(\beta)$.

(b) There exists some $\beta \in \alpha$ such that $\text{succ}(\beta)$ does not belong to $\alpha$.

(c) We have $\bigcup \alpha \in \alpha$.

If $\alpha$ is neither zero nor a successor ordinal, we call it a **limit ordinal**. See proposition 1252 for a more involved equivalent condition.

These notions should not be confused with successor and weak/strong limit cardinals.

**Proof.**

**Proof that 1023 (a) implies 1023 (b).** If $\alpha = \text{succ}(\beta)$, then $\beta$ satisfies definition 1023 (b).

**Proof that 1023 (b) implies 1023 (a).** Let $\beta \in \alpha$ be such that $\text{succ}(\beta) \notin \alpha$. Then by trichotomy we have that either $\text{succ}(\beta) = \alpha$ or $\text{succ}(\beta) > \alpha$.

If $\text{succ}(\beta) > \alpha$, then either $\alpha = \beta$, which would contradict corollary 994 (a), or $\alpha \in \beta$, which would contradict corollary 994 (b).

Thus, it remains for $\text{succ}(\beta)$ to be equal to $\alpha$.

**Proof that 1023 (a) implies 1023 (c).** Suppose that $\alpha = \text{succ}(\beta)$.

We have

\[
\bigcup \alpha = \bigcup (\beta \cup \{\beta\}) = \\
= \{\gamma \mid \exists \delta \cdot (\delta \in \beta \text{ or } \delta = \beta) \text{ and } \gamma \in \delta\} \overset{(595)}{=} \\
= \{\gamma \mid \exists \delta \cdot (\delta \in \beta \text{ and } \gamma \in \delta) \text{ or } (\delta = \beta \text{ and } \gamma \in \delta) = \\
= \{\gamma \mid (\exists \delta \in \beta \cdot \gamma \in \delta) \text{ or } \gamma \in \beta\} = \\
= (\bigcup \beta) \cup \beta \overset{\text{UPC}}{=} \\
= \beta.
\]

Thus, $\beta = \bigcup \alpha \in \alpha$.

**Proof that 1023 (c) implies 1023 (a).** Let $\bigcup \alpha \in \alpha$. From remark 1022 we have that

\[
\text{succ}(\bigcup \alpha) = \{\gamma \mid \gamma \subseteq \bigcup \alpha\} = \{\gamma \mid \gamma \not\subseteq \alpha\} = \alpha.
\]

**Remark 1024.** It is sometimes efficient to reformulate transfinite induction and recursion. Although analogous principles hold for theorem 1009 (Bounded transfinite induction) and theorem 1010 (Bounded transfinite recursion), we will only demonstrate them for theorem 1011 (Transfinite induction). The original statement is that in order to prove that some formula is satisfied for all sets, it is sufficient to only prove one inductive step.
More precisely, let $\varphi$ be a formula in the language of set theory, let $\mathcal{V} = (V, I)$ be a standard transitive model of set theory and let $v : \text{Var} \to V$ be some variable assignment. This allows us to fix any parameters that would otherwise be present in the induction schema. We will say that the set $A$ satisfies $\varphi$ if $[v_{\varphi_A}] = T$.

Theorem 1011 (Transfinite induction) states that in order to prove that $\varphi$ holds for any ordinal, the following is sufficient:

(a) For every ordinal $\alpha$, by assuming that every smaller ordinal satisfies $\varphi$, we must prove that $\alpha$ does.

We have just defined in definition 1023 three mutually exclusive types of ordinals. We can now restate the principles of transfinite induction as follows:

(b) In the base case, we must prove that $0$ satisfies $\varphi$.

(c) In the successor case, by assuming that $\alpha$ satisfies $\varphi$, we must prove that $\text{succ}(\alpha)$ does. It is in line with theorem 1011 (Transfinite induction) assume that every ordinal smaller that or equal to $\alpha$ satisfies $\varphi$, however it is often enough to do, so only for $\alpha$ itself.

(d) In the limit case, if $\lambda$ is a limit ordinal, by assuming that every smaller ordinal satisfies $\varphi$ we must prove that $\lambda$ satisfies $\varphi$.

We must note that remark 1024 (d) is essentially the same as the single inductive step remark 1024 (a) except that it is restricted to limit ordinals. The reasoning for this is that the proofs for zero and every successor ordinal can be different from those for limit ordinals. See lemma 1030 and definition 1091 for concrete examples.

Theorem 1025 (Structural recursion). The most general recursion principle we will consider is structural recursion. It is sometimes also called mutual recursion. See theorem 997 (Well-founded induction) for its corresponding induction principle.

Let $X$ be a set and let $T : \text{pow}(X) \to \text{pow}(X)$ be a transformation. Then $T$ has a smallest fixed point. That is, there exists a unique subset $A_0 \subseteq X$ such that $T(A_0) = A_0$ and no proper subset of $A_0$ has this property.

See remark 1026 for how this theorem is used.

Proof. The result follows by applying theorem 1259 (Knaster-Tarski theorem) to the Boolean algebra of all subsets of $X$ with $R(A) := A \cup T(A)$ as the monotone operator. □

Remark 1026. We will now show the connection between theorem 997 (Well-founded induction) and theorem 1025 (Structural recursion).

Let $\mathcal{F}$ be a set of functions, where each function $f$ has a signature $f : X^{#f} \to X$ for some nonnegative integer $#f$. Define the operator

$$T : \text{pow}(X) \to \text{pow}(X)$$

$$T(A) := \left\{ x \in X \mid \exists f \in \mathcal{F} : \exists t_1, \ldots, t_{#f} \in A_0 . f(t_1, \ldots, t_{#f}) = x \right\}.$$
Now we can use theorem 1025 (Structural recursion) to obtain the smallest fixed point \( A_0 \) of \( T \). The set \( A_0 \) is closed under all the functions in \( \mathcal{F} \), i.e. \( f[A_0^f] \subseteq A_0 \) for any \( f \in \mathcal{F} \).

Define a binary relation \( \rightarrow \) on \( A_0 \) by declaring that, for every function \( f \) in \( \mathcal{F} \), every sequence \( x_1, \ldots, x_{\#f} \) and every index \( k = 1, \ldots, \#f \), we have

\[
x_k \rightarrow f(x_1, \ldots, x_{\#f}).
\]

This relation can be defined by taking unions of smaller relations rather than via recursion. If the quiver happens to be well-founded, we can use theorem 997 (Well-founded induction) to prove universal statements about \( A_0 \).

In section 12.5 (First-order satisfiability), for example, we use structural recursion to define first-order substitution in definition 835 and then use structural induction to prove certain semantic equivalences like proposition 839 and proposition 840.

Remark 1027. **Mathematical induction** is a very valuable proof technique for universal statements. The proof of proposition 5 contains remarks regarding its usage and its difference from deduction principles that are formalized via deductive systems. Although it is a logical tool, this remark belongs to this section because it contains several induction and recursion principles.

More generally, given a first-order formula \( \varphi \) over some first-order language, certain logical theories allow us to prove indirectly \( \forall \xi. \varphi[\xi \mapsto \eta] \) by proving simpler statements (definition 1 contains very important remarks regarding the free variables of \( \varphi \)). This can be done in cases where every model \( X = (X, I) \) of the theory allows us to exhaust its universe \( X \) in a small finite number of steps. We can sometimes use the same steps to instead build objects. The latter principles is called recursion.

It should be noted that induction and recursion are used interchangeably in the literature, especially regarding structural induction, however we will aim to distinguish between the two.

Not much more can be said at this level of generality, so we list several induction principles and give examples of their usage:

(a) The most basic induction principles is the (weak) natural number induction. It is best described via the axiom schema (PA3). Proposition 5 contains detailed commentary regarding its usage and most of the proofs in section 1.1 (Natural numbers) are performed inductively.

It its set-theoretic form theorem 943 (Induction via inductive sets) it is important as a tool for introducing a model of Peano arithmetic. It is used directly for proving proposition 945 and theorem 979 (Recursion theorem). It is essentially the same as remark 1024 without remark 1024 (d).

Theorem 979 (Recursion theorem) is an important standalone tool that allows us to perform recursive definitions for natural numbers. The latter is used, often implicitly, in a great variety of places, from the ability to define natural number operations in definition 980 to the definition of magma exponentiation in definition 441 (e). Remark 982 contains notes regarding its practical usage.
(b) A vast generalization of natural number induction is theorem 997 (Well-founded induction). It is stated in a very general setting, but is not frequently used. It can be used to prove theorem 998 (Epsilon induction), which however is even less frequently used. We do not use neither in practice, however the special case where $X = \mathbb{N}$ is called strong induction on natural numbers.

The usual (weak) natural number induction which is performed by proving the statement for $0$ and then proceeding to prove it for $n + 1$ by assuming that it holds for $n$. Strong induction instead has no base cases and is performed by proving a statement for $n$ by assuming that it holds for all natural numbers strictly smaller than $n$. Well-founded induction and epsilon-induction have no corresponding recursion principle.

(c) Another vast generalization of natural number induction is theorem 1009 (Bounded transfinite induction), which is further generalized by theorem 1011 (Transfinite induction).

Both principles are used to prove fundamental properties of the ordinals. Outside of set theory, transfinite induction is immensely useful, but it is rarely used directly. Instead, it is usually combined with the axiom of choice via theorem 1240 (Zorn’s lemma). Remark 1024 contains notes on how it is used directly.

Only bounded transfinite induction has a corresponding recursion principle — theorem 1010 (Bounded transfinite recursion). Unbounded transfinite induction cannot define such a principle because that would easily lead to theorem 1019 (Burali-Forti paradox). Remark 1054 (Unbounded transfinite recursion) shows how to circumvent this, however.

Transfinite recursion is used to construct the cumulative hierarchy in definition 1091. Remark 1056 (Cardinal recursion and induction) provides alternative transfinite recursion and induction principles for cardinals rather than for ordinals.

(d) A very general recursion principle is theorem 1025 (Structural recursion). Inductive proofs can be performed via theorem 997 (Well-founded induction) sometimes — this is discussed in remark 1026.

Within this document, we use theorem 792 (Structural induction on unambiguous grammars) in certain special cases, for example propositional and first-order formulas.

**Proposition 1028.** Two ordinals are equal if and only if they are order-isomorphic.

**Proof.**

**Proof of sufficiency.** Trivial.

**Proof of necessity.** Let $\alpha$ and $\beta$ be two ordinals. The case $\alpha = \beta$ is clear. Without loss of generality, suppose that $\beta < \alpha$. Proposition 1013 implies that $\beta \subsetneq \alpha$.

Let $f : \alpha \to \beta$ be an order isomorphism. Let $\gamma_0$ be the smallest value in $\alpha \setminus \beta$.

From lemma 999 it follows that $\gamma_0 \leq f(\gamma_0)$. But $f(\gamma_0) \in \beta$, hence $\gamma_0 < \beta$. But this contradicts our choice of $\gamma_0$ as a member of $\alpha \setminus \beta$. 

524
Therefore, \( \gamma_0 \in \beta \). Since \( \gamma_0 \) was arbitrary, we conclude that \( \beta < \alpha \) leads to a contradiction.

We can obtain an analogous contradiction for \( \alpha < \beta \), hence it only remains possible for \( \alpha \) and \( \beta \) to be equal. \( \square \)

**Theorem 1029** (Hartogs’ lemma). *For every set \( A \) there exists a smallest ordinal \( \alpha \) such that no function from \( \alpha \) to \( A \) is injective.*

*Proof.* Define the set

\[
W := \{(P, \leq) \mid P \subseteq A \text{ and } \leq \text{ well-orders } P\}.
\]

Let \( \beta \) be an ordinal for which there exists an injective function \( f : \beta \to A \). At least one such pair of a ordinal and function exists because \( \emptyset : 0 \to A \) is an injective function.

The image of \( f \) can be well-ordered by

\[
f(\delta) < f(\gamma) \text{ if and only if } \delta \in \gamma,
\]

where \( \delta \) and \( \gamma \) are members of \( \beta \). Thus, if we restrict the range of \( f \) to its image, it would be an explicit order isomorphism of \((\beta, \in)\) and \((\text{img } f, <)\).

We have shown that every ordinal \( \beta \) and every injective function \( f : \beta \to A \) induces a well-ordered set which belongs to \( W \). Furthermore, if \( f_1 : \beta_1 \to A \) and \( f_2 : \beta_2 \to A \) are two such injective functions and if the induced well-ordered sets \((\text{img } f_1, \leq_1)\) and \((\text{img } f_2, \leq_2)\) are order-isomorphic, then \( \beta_1 \) and \( \beta_2 \) are themselves order-isomorphic and thus \( \beta_1 = \beta_2 \) by proposition 1028.

Therefore, for any well-ordered set in \( W \) there exists at most one ordinal that induces it via some function. Let \( W' \) be the subset of all well-ordered sets in \( W \) induced by exactly one ordinal.

We have that \( W' \) is a set and for each member of \( W' \) there corresponds exactly one ordinal. It follows from the axiom schema of replacement that these ordinals form a set. Denote this set by \( B \).

There is no set of all ordinals by theorem 1019 (Burali-Forti paradox), hence there must exist an ordinal not in \( B \). But every ordinal that has an injective function into \( A \) belongs to \( B \). Hence, there exists some ordinal \( \alpha \) such that no function from \( \alpha \) to \( A \) is injective.

If \( \alpha \) is not the smallest ordinal with this property, we can now easily take the smallest member of \( \alpha \) with this property. \( \square \)

**Lemma 1030.** Let \( A \) be a nonempty set and let \( c \) be a choice function on \( A \). Then there exists an ordinal \( \alpha \) and a bijection between \( A \) and \( \alpha \).

*Proof.* We will explicitly build the desired ordinal. We will use theorem 1010 (Bounded transfinite recursion) in its structured form presented in remark 1024 to build a transfinite sequence of injective maps from ordinals into \( A \).

Let \( \delta \) be the smallest ordinal such that no function from \( \delta \) to \( A \) is injective. Such an ordinal exists by theorem 1029 (Hartogs’ lemma). Note that \( \delta \) cannot be zero because the empty function is always injective.

Any smaller than \( \delta \) ordinal has at least one injective function into \( A \). We will use recursion bounded by \( \delta \) to construct a transfinite sequence \( \{i_\beta\}_{\beta < \delta} \) satisfying the invariant that for any \( \beta \in \delta \), the function \( i_\beta \) is injective and strictly contains \( i_\gamma \) for any \( \gamma < \beta \).
• The zero ordinal has only one possible injective function into \( A \), namely the empty function \( \emptyset : 0 \to P \).

• Now let \( t_\beta : \beta \to A \) be an injective function.

  If \( t_\beta \) is surjective, then it is the desired bijection and the rest of the construction is irrelevant.

  If \( t_\beta \) is not surjective, define

  \[
  t_{\text{succ}(\beta)} : \text{succ}(\beta) \to P
  \]

  \[
  t_{\text{succ}(\beta)}(\gamma) := \begin{cases} t_\beta(\gamma), & \gamma < \beta \\ c(A \setminus \text{img}(t_\beta)), & \gamma = \beta. \end{cases}
  \]

  This function is clearly injective.

• Let \( \lambda \) be a limit ordinal and let \( t_\beta \) be an injective function for any \( \beta < \lambda \). Note that the domain of \( t_\beta \) is not necessarily \( \beta \) — it may be a smaller ordinal \( \gamma \) in case \( t_\gamma \) is surjective (and thus the desired bijection).

  In any case, we have a family \( \{t_\beta\}_{\beta < \lambda} \) of functions such that \( t_\gamma \subseteq t_\beta \) if and only if \( \gamma < \beta \).

  We simply take their union

  \[
  t_\lambda := \bigcup \{t_\beta \mid \beta < \lambda\}.
  \]

  This function is injective because each of the constituent functions it injective.

  We will now thoroughly build the order isomorphism itself.

  Let \( \alpha \leq \delta \) be the (nonstrict) initial segment of \( \delta \) such that \( t_\beta \) is fully defined on \( \beta \) for every \( \beta \in \alpha \). More precisely, let

  \[
  \alpha := \{\beta \in \delta \mid \text{dom}(t_\beta) = \beta\}.
  \]

  We will show that \( \alpha \) is a successor ordinal. Note that \( \alpha \) cannot be zero because \( \text{dom}(t_0) = \emptyset = 0 \). Aiming at a contradiction, assume that \( \alpha \) is a limit ordinal.

  Since for every \( \beta \in \alpha \) the value \( t_{\text{succ}(\beta)}(\beta) \) is defined, we can conclude that the union

  \[
  \bigcup \{t_\beta \mid \beta \in \alpha\}
  \]

  is an injective function from \( \alpha \) to \( A \).

  If \( \alpha = \delta \), this contradicts our choice of \( \delta \). If \( \alpha \in \delta \), this means that the function \( t_\xi \) is equal to \( t_\gamma \) for some \( \gamma < \alpha \). But then \( t_{\text{succ}(\gamma)} \) is not defined on the entirety of \( \gamma \). This implies that \( \gamma \geq \alpha \), which contradicts our assumption that \( \gamma < \alpha \).

  Therefore, \( \alpha \) cannot be a limit ordinal. It remains for \( \alpha \) to be a successor ordinal. Then there exists \( \beta \) such that \( \alpha = \text{succ}(\beta) \).

  Suppose that \( t_\beta \) is not surjective. We can then construct \( t_{\text{succ}(\beta)} \) as in the corresponding recursion step. This will be an injective function from \( \alpha = \text{succ}(\beta) \) to \( A \). But then we would have \( \alpha \in \alpha \), which contradicts corollary 994 (a).

  Therefore, \( t_\beta \) is a surjective function. \( \square \)
Theorem 1031. Any well-ordered set \((P, \leq)\) is order-isomorphic to a unique ordinal. This ordinal is called the order type of \((P, \leq)\) and is denoted by \(\text{ord}(P, \leq)\) or simply \(\text{ord}(P)\). Furthermore, this isomorphism is unique.

Proof. For \(P = \emptyset\), then the empty function \(\iota : 0 \rightarrow P\) is the desired isomorphism.

We use lemma 1030 on \(P\) with the choice function

\[
c : \text{pow}(P) \setminus \{\emptyset\} \rightarrow P
\]

\[
c(B) := \min B
\]

to obtain an ordinal \(\alpha\) and a bijection \(\iota : \alpha \rightarrow P\).

The function \(\iota\) is directly verified to be an order embedding in each of the recursion cases. Therefore, \(\iota\) is a bijective order homomorphism and it follows from corollary 1244 that \(\iota\) is an order isomorphism.

Uniqueness of \(\iota\) follows from proposition 1000. \hfill \Box

Theorem 1032 (Zermelo’s well-ordering theorem). Any set can be well-ordered. Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (j).

Proof.

Proof that the axiom of choice implies well-ordering theorem. The empty set is trivially well-ordered.

Let \(A\) be any nonempty set. By the axiom of choice, there exists a choice function \(c\) for \(A\). We use lemma 1030 on \(A\) and \(c\) to obtain an ordinal \(\alpha\) and a bijection \(\iota : \alpha \rightarrow A\). Since \(\alpha\) itself is well-ordered, then the relation

\[
x <_A y \text{ if and only if } r^{-1}(x) <_\alpha r^{-1}(y)
\]

defines a well-order on \(A\).

Proof that well-ordering theorem implies axiom of choice. Let \(A\) be any set and suppose that \(\prec\) well-orders \(A\).

Define the function

\[
c : \text{pow}(A) \setminus \{\emptyset\} \rightarrow A
\]

\[
c(B) := \min B.
\]

In the case where \(A\) is the empty set, \(c\) is the empty function.
It is clear that \(c(B) \in B\) for every subset \(B\) of \(A\). Therefore, \(c\) is a choice function for \(A\). \hfill \Box
13.7. Cardinals

**Definition 1033.** We say that two sets are **equinumerous** if there exists a bijective function between them.

If there exists an injective function from $A$ to $B$ that is not necessarily surjective, we say that $A$ is **dominated by** $B$ or that $B$ **dominates** $A$. If $B$ dominates $A$, and they are not equinumerous, we say that $B$ **strictly dominates** $A$.

Equinumerosity arises naturally outside the theory of cardinal numbers, unlike set dominance. We are usually instead interested only in injective functions that preserve some structure, i.e. embeddings.

**Lemma 1034.** If $A \subseteq B \subseteq C$ and $A$ is equinumerous with $C$, then $B$ is equinumerous with $C$.

**Proof.** If $B = C$, the lemma is trivial since the identity function $\text{id}_B : B \to B$ is bijective. If $B \nsubseteq C$, however, the identity $\text{id}_B$ must be extended in order to be a bijective function between $B$ and $C$. It will actually be simpler for us to define a function from $C$ to $B$.

Let $f : C \to A$ be a bijective function (such a function exists by the statement of the lemma). Define the set

$$ I := \bigcap \{X \subseteq C \mid (C \setminus B) \subseteq X \text{ and } f(X) \subseteq X \} $$

of all intermediate sets between $C \setminus B$ and $C$ that are invariant under $f$.

Use natural number recursion to build the function

$$ g : C \to B $$

$$ g(x) := \begin{cases} x, & x \in C \setminus I \\ f(x), & x \in I \end{cases} $$

By construction, $C \setminus B \subseteq I$ and thus $C \setminus I \subseteq C \setminus (C \setminus B) = B$. Therefore, the range of $g$ really is $B$. We must show that $g$ is injective and surjective.

Let $g(x_1) = g(x_2)$ for some members $x_1$ and $x_2$ of $C$. If $x_1$ and $x_2$ both belong to either $I$ or $C \setminus I$, it is trivial to see that $x_1 = x_2$. It turns out that these are two only possible scenarios. Indeed, without loss of generality, suppose that $x_1 \in I$ and $x_2 \in C \setminus I$. Then $f(x_2) = g(x_2) = g(x_1) = x_1$. Since $I$ is invariant under $f$ and $x_1 \in I$, we have $x_2 = f(x_1) \in I$, which contradicts our choice of $x_2$. Therefore, $g$ is an injective function.

To see that $g$ is also surjective, suppose that there exists some $y \in B \setminus g[B]$. If $y \in I$, then by the invariance of $f$ we have $g(y) = f(y) \in I$ and thus $g(y) \notin B$, which contradicts our definition of $g$. If instead $y \in C \setminus I$, then $g(y) = y$ and thus $y \in g[B]$, which contradicts our choice of $y$. The obtained contradictions show that $g$ is surjective.

**Theorem 1035 (Cantor-Schröder-Bernstein theorem).** If two sets dominate each other, they are equinumerous.

**Proof.** Let $f : A \to B$ and $g : B \to A$ be injective functions. From proposition 972 it follows that $g \circ f : A \to A$ is also an injective function. If we restrict its range to its image $g[f[A]]$, it becomes bijective. Hence, $A$ is equinumerous with $g[f[A]]$. Since $g[f[A]] \subseteq g[B] \subseteq A$, from lemma 1034 it follows that $A$ is equinumerous with $g[B]$, which is the desired result.
Set domination generalizes the subset relation between sets. If we take a family $\mathcal{A}$ of sets, then domination is a preorder rather than a true partial order.

- Reflexivity follows because the identity function for any set is injective.
- Transitivity is a consequence of proposition 972.
- Antisymmetry fails if $A$ dominates $B$ and $B$ dominates $A$, but $A \neq B$. For example, the map $n \mapsto 2n$ from all natural numbers $\mathbb{N}$ to the even natural numbers $2\mathbb{N}$ is injective and the identity map on $2\mathbb{N}$ is an injective function from $2\mathbb{N}$ to $\mathbb{N}$, however $\mathbb{N} \neq 2\mathbb{N}$.

As a matter of fact, equinumerosity is also a preorder and the proof for that is identical. If we partition $\mathcal{A}$ using the equinumerosity relation if follows from proposition 1226 that the result will be a partial ordered set. The equivalence classes of this partition are then subfamilies of $\mathcal{A}$ such that every two sets in a single subfamily are equinumerous. For example, if $\mathcal{A} = \{\{A, B\}, \{C, I\}, \{A\}, \{C\}\}$, then the corresponding equivalence classes are $\{\{A, B\}, \{C, I\}\}$ and $\{\{A\}, \{C\}\}$.

Each of these equivalence classes consists of sets that are identical in “size” (not to be confused with “large” and “small” sets as defined in definition 1112). In the above example, the corresponding equivalence classes correspond to sets of sizes 1 and 2. If we want to extend this notion of “size” to infinite sets, we must introduce a hierarchy of “sizes”. A natural candidate for such a hierarchy are the equivalence classes themselves. Unfortunately, this would mean that every family of sets has a different hierarchy. Since the entire universe is only available within the metatheory, we cannot partition the universe itself and must instead resort to finding a concrete representative of each possible equivalence class. We will call these representatives cardinal numbers.

As explained in the proof of proposition 1041, it will be convenient for us to define cardinal numbers as certain ordinal numbers — see definition 1037.

**Definition 1037.** A cardinal number or simply cardinal is an ordinal that is not equinumerous with any smaller ordinal. We usually denote them using the small Greek letters $\kappa$, $\mu$ and $\nu$.

A cardinal is by definition an ordinal and this is useful. For example, the cardinals are well-ordered in the sense of proposition 1012.

We often regard cardinal numbers as abstract entities, however. It is thus accepted to call the ordinal itself the initial ordinal of the cardinal.

**Lemma 1038.** No natural number (as a member of $\omega$) is equinumerous to a proper subset of itself.

**Proof.** We will use theorem 943 (Induction via inductive sets) on $n \in \omega$. The lemma holds vacuously for $n = 0$.

Now suppose that $n$ is not equinumerous to a proper subset of itself and, aiming at a contradiction, suppose that there exists a subset $E \subseteq \text{succ}(n)$ and a bijective function $f : \text{succ}(n) \to E$.

- If $n \in E$, then $E \setminus \{n\}$ is a subset of $n$ and thus the restriction $f|_n : n \to (E \setminus \{n\})$ is a bijective function.
If \( n \notin E \), then \( E \) is a subset of \( n \) and thus \( f|_n : n \to E \) is a bijective function.

In both cases we obtain that \( n \) is equinumerous with a proper subset of itself, which is a contradiction. Hence, this also holds for \( \text{succ}(n) \).

The induction principle allows us to conclude that the lemma holds for all natural numbers. \( \square \)

**Proposition 1039.** The natural numbers (as members of \( \omega \)) are cardinals.

**Proof.** Fix a natural number \( n \in \omega \). Note that for every \( m < n \), \( m \) is a proper subset of \( n \) by proposition 1013. From lemma 1038 it follows that no ordinal strictly smaller than \( n \) is equinumerous with \( n \) and hence \( n \) is an initial ordinal. \( \square \)

**Proposition 1040.** The smallest inductive set \( \omega \) is a cardinal.

When regarded as a cardinal, we denote it by \( \aleph_0 \). This is consistent with definition 1053.

**Proof.** We will use induction on \( n < \omega \) to show that no function \( f : \omega \to n \) is surjective. This is trivial for 0. Suppose that it holds for some fixed \( n \). Let \( f : \omega \to \text{succ}(n) \) be any function. Define

\[
g : \omega \to n, \quad g(m) := \begin{cases} f(m), & f(m) < n \\ 0, & f(m) = n \end{cases}
\]

Our inductive hypothesis states that \( g \) cannot be surjective. Hence, \( f \) also cannot be surjective.

Therefore, \( \omega \) is an initial ordinal. \( \square \)

**Proposition 1041.** Every set \( A \) is equinumerous with a unique cardinal. We denote this cardinal by \( \text{card}(A) \) and call it the cardinality of \( A \).

**Proof.** By theorem 1032 (Zermelo’s well-ordering theorem) there exists a relation \( < \) that well-orders \( A \). The order type \( \text{ord}(A, <) \) is an ordinal that is equinumerous with \( A \), however it may not be the smallest one. Fortunately, we can define

\[
\text{card}(A) := \min\{\beta \leq \text{ord}(A, <) \mid \beta \text{ is equinumerous with } A\}.
\]

\( \square \)

**Proposition 1042.** The set \( A \) is dominated by \( B \) if and only if \( \text{card}(A) \leq \text{card}(B) \).

**Proof.**

**Proof of sufficiency.** First suppose that \( \text{card}(A) \leq \text{card}(B) \). By proposition 1013, we have \( \text{card}(A) \subseteq \text{card}(B) \) and thus the identity function \( \text{id}_{\text{card}(A)} \) is an injective function from \( \text{card}(A) \) to \( \text{card}(B) \). Since \( A \) is equinumerous with \( \text{card}(A) \) and \( B \) is equinumerous with \( \text{card}(B) \), by proposition 972 we obtain that there is an injective function from \( A \) to \( B \) and hence \( B \) dominates \( A \).
Proof of necessity. Conversely, let \( f : A \to B \) be an injective function. We again use proposition 972 to conclude that \( \text{card}(B) \) dominates \( \text{card}(A) \).

We will show that \( \text{card}(A) > \text{card}(B) \) leads to a contradiction, which by the trichotomy of cardinals will entail that \( \text{card}(A) \leq \text{card}(B) \). If we suppose that \( \text{card}(A) > \text{card}(B) \), then \( \text{card}(B) \subseteq \text{card}(A) \) and hence the identity on \( \text{card}(B) \) is an injective function. Thus, \( \text{card}(A) \) dominates \( \text{card}(B) \) and vice versa, which by theorem 1035 (Cantor-Schröder-Bernstein theorem) implies that \( \text{card}(A) \) is equinumerous with \( \text{card}(B) \). It follows that \( \text{card}(A) = \text{card}(B) \), which contradicts our assumption that \( \text{card}(A) > \text{card}(B) \).

Therefore, \( \text{card}(A) \leq \text{card}(B) \). \( \square \)

Corollary 1043. Any two sets are either equinumerous or one strictly dominates the other.

See also theorem 1284 (Dirichlet’s pigeonhole principle).

Proof. Follows from cardinal trichotomy and proposition 1042. \( \square \)

Theorem 1044 (Cantor’s power set theorem). The power set of any set \( A \) strictly dominates \( A \).

That is,

\[
\text{card}(A) < \text{card}(\text{pow}(A)).
\]

Proof. The function \( x \mapsto \{x\} \) is clearly an injective function from \( A \) to \( \text{pow}(A) \), therefore \( \text{pow}(A) \) dominates \( A \). The converse is not true, however.

Indeed, fix some function \( f : A \to \text{pow}(A) \) and define the set

\[
B := \{x \in A : x \notin f(x)\}.
\]

Note that \( B \subseteq A \) and thus \( B \in \text{pow}(A) \), however \( B \) is not in the image of \( f \) and thus \( f \) is not surjective.

Since \( f \) was arbitrary, we conclude that no function from \( A \) to \( \text{pow}(A) \) is surjective. \( \square \)

Definition 1045. We say that the set \( A \) is finite if any of the following equivalent conditions hold:

(a) The cardinality of \( A \) is a natural number. That is, we have \( \text{card}(A) < \aleph_0 \).

(b) The set \( A \) is not equinumerous all of its proper subsets. That is, if \( B \) is a proper subset of \( A \), then no function from \( A \) to \( B \) is injective.

If a set is not finite, we say that it is infinite. If a set does not satisfy definition 1045 (b), we say that it is Dedekind infinite.

Proof.

Proof that 1045 (a) implies 1045 (b). We will use theorem 943 (Induction via inductive sets) on \( n < \aleph_0 \) to prove that all sets of cardinality \( n \) are not Dedekind infinite. The case \( n = 0 \) is vacuous. Suppose that all sets of cardinality \( n \) are not Dedekind infinite and suppose that \( A \) of cardinality \( \text{succ}(n) \) is Dedekind infinite.

Then there exists a proper subset \( B \) of \( A \) that is equinumerous with \( \text{succ}(n) \). Since \( \text{succ}(n) \) is the cardinality of \( A \), there exists a bijective function \( f : A \to \text{succ}(n) \). Then \( f[B] \subseteq \text{succ}(n) \).
Furthermore, the inequality is strict because otherwise any member of \( A \setminus B \) would contradict theorem 1284 (Dirichlet’s pigeonhole principle). Therefore, \( f[B] \) is a proper subset of \( \text{succ}(n) \) that is equinumerous with it. But this contradicts lemma 1038.

The obtained contradiction shows that \( A \) is also not Dedekind infinite.

**Proof that 1045 (b) implies 1045 (a).** Let \( A \) be a Dedekind infinite set and let \( f : A \to B \) be a bijective function into some proper subset \( B \) of \( A \). We will construct an injective function from \( \omega \) to \( A \). Fix some member \( x_0 \in A \setminus B \) and recursively define

\[
g : \omega \to A \quad g(n) := \begin{cases} x_0, & n = 0 \\ f(g(n - 1)), & n > 0 \end{cases}
\]

We now use induction on \( m \) to prove that \( g(n) = g(m) \) implies \( n = m \). Since \( g(0) \notin B \) and \( g(n) \in B \) for any \( n > 0 \), the base case holds. Suppose that the inductive hypothesis holds for \( m \) and that for some \( n \) we have \( g(n) = g(m + 1) \). It is clear that \( g(0) \neq g(m + 1) \), so necessarily \( n > 0 \). We have

\[
f(g(n - 1)) = g(n) = g(m + 1) = f(g(m)),
\]

which by the inductive hypothesis implies that \( m = n - 1 \). Thus, \( g(m + 1) = g(n) \) and the inductive step is proved.

Therefore, \( g \) is injective and thus \( A \) dominates \( \omega \). From proposition 1042 it follows that \( \text{card}(A) \geq \aleph_0 \) and thus \( \text{card}(A) \) is not a natural number.

**Proposition 1046.** A nonzero cardinal is finite if and only if it is a successor ordinal.

**Proof.** Finite cardinals are natural numbers by definition and all nonzero natural numbers are successor ordinals.

Conversely, suppose that \( \kappa = \text{succ}(\alpha) \) is a successor ordinal that is a cardinal. That is, \( \kappa \) is not equinumerous with any smaller ordinal and in particular with \( \alpha \). From proposition 1042 it follows that \( \text{card}(\alpha) < \kappa \).

Let \( A \subseteq \kappa \) be a proper subset of \( \kappa \). We want to show that \( \text{card}(A) < \kappa \).

If \( \alpha \notin A \), define \( B := A \). Otherwise, pick some member \( x_0 \in \kappa \setminus A \) and define

\[
B := (A \cup \{x_0\}) \setminus \{\alpha\}.
\]

In both cases we have \( \text{card}(A) = \text{card}(B) \), but unlike \( A \), \( B \) is always a subset of \( \alpha \) since \( \kappa = \alpha \cup \{\alpha\} \).

Therefore,

\[
\text{card}(A) = \text{card}(B) \leq \text{card}(\alpha) < \kappa.
\]

Hence, \( \kappa \) dominates every proper subset, which by definition 1045 (b) means that \( \kappa \) is a finite cardinal.

**Corollary 1047.** A cardinal is infinite if and only if it is a limit ordinal.

**Proof.** This is the contraposition to proposition 1046 excluding the zero cardinal.
**Proposition 1048.** A set is finite if and only if its power set is finite.

**Proposition 1049.** All finite unions and Cartesian products of finite sets are finite.

**Definition 1050.**
(a) If $\kappa$ is the smallest cardinal such that $\mu < \kappa$ for some other cardinal $\mu$, we say that $\kappa$ is the **successor** of $\mu$ and that $\kappa$ is itself a **successor cardinal**.

The existence of $\kappa$ is guaranteed by **proposition 1051**, but it is natural to ask whether $\kappa$ can be constructed from $\mu$ similarly to how the ordinal successor operator gives us a successor ordinal. This turns out to be a deep question — see **Conjecture 1088** (Generalized continuum hypothesis).

(b) If $\kappa > 0$ is not the successor cardinal of any other cardinal, we say that it is a **weak limit cardinal**.

See **corollary 1058** for some equivalent conditions.

(c) We say that $\kappa$ is a **strong limit cardinal** if $\mu < \kappa$ implies that $\text{card}(\text{pow}(\mu)) < \kappa$.

We can benefit from using forward references to **section 13.8** (Transfinite arithmetic), more precisely **proposition 1086**, which justifies using cardinal exponentiation to rewrite the condition for $\kappa$ being a strong limit cardinal as $\mu < \kappa$ implies $2^\mu < \kappa$.

Every strong limit cardinal is a weak limit cardinal as shown in **proposition 1052**, however the converse is only true assuming **Conjecture 1088** (Generalized continuum hypothesis) — see **corollary 1089**.

Strong limit cardinals are further motivated by the usage of regular strong limit cardinals in **proposition 1102**.

These notions should not be confused with successor and limit ordinals.

**Proposition 1051.** For any cardinal there exists a successor cardinal.

**Proof.** Fix a cardinal $\kappa$. By **theorem 1029** (Hartogs’ lemma), there exists a smallest ordinal $\alpha$ such that $\kappa$ does not dominate $\alpha$. Thus, $\alpha$ is the initial ordinal of a cardinal $\mu$ because it is not equinumerous with any smaller ordinal.

**Corollary 1043** implies that $\kappa < \mu$.

Furthermore, every cardinal smaller than $\mu$ does not dominate $\kappa$, i.e. if $\nu < \mu$, then $\nu \geq \kappa$.

Therefore, $\mu$ is the successor cardinal of $\kappa$. $\Box$

**Proposition 1052.** If $A$ is a set of cardinals, then $\bigcup A$ is a cardinal. Furthermore, $\bigcup A$ is the supremum of $A$ with respect to cardinal ordering.

**See a more thorough discussion of a similar issue in proposition 1020.**

**Proof.** From **proposition 1020** it follows that $\bigcup A$ is an ordinal. Then there exists some cardinal $\kappa \in A$ such that $\alpha \in \kappa$.

We have $\kappa \subseteq \bigcup A$. Thus, with regards to ordinal ordering, $\alpha \prec \kappa \subseteq \bigcup A$. But since $\kappa$ is a cardinal, it is not equinumerous with $\alpha$ and hence $\bigcup A$ is also not equinumerous with $\alpha$.

Therefore, $\bigcup A$ is a cardinal. It follows from **proposition 1013** that it is also the supremum of $A$. $\Box$
Definition 1053. We use transfinite recursion to define, for each ordinal \( \alpha \), the cardinal

\[
\kappa_\alpha := \begin{cases} 
\omega, & \alpha = 0 \\
\text{successor cardinal of } \beta, & \alpha = \text{succ}(\beta) \\
\sup\{\kappa_\beta \mid \beta < \alpha\} & \alpha \text{ is a limit ordinal}
\end{cases}
\]

(355)

We denote the initial ordinal of \( \kappa_\alpha \) by \( \omega_\alpha \). In particular, \( \omega_0 = \omega \) and \( \omega_1 \) is the first uncountable ordinal.

Note that \( \kappa_\lambda \) exists and is a cardinal for every limit ordinal \( \lambda \) as a consequence of proposition 1052.

See remark 1054 (Unbounded transfinite recursion) for some technical details.

This hierarchy is important because it describes all infinite cardinals as shown in proposition 1057. It is intimately connected to the simpler \( \beth \) hierarchy via Conjecture 1088 (Generalized continuum hypothesis).

Remark 1054 (Unbounded transfinite recursion). Although we cannot formally do unbounded transfinite recursion, there is an easy way to circumvent this.

Formally, in definition 1053, for every ordinal \( \alpha \) we use theorem 1010 (Bounded transfinite recursion) define a \( \alpha \)-indexed transfinite sequence \( \kappa_0, \kappa_1, \ldots, \kappa_\omega, \ldots \) and then use the sequence to define \( \kappa_\alpha \). The definition does not depend on any particular ordinal \( \alpha \), however, and thus all ways to obtain \( \kappa_\alpha \) are equivalent.

Proposition 1055. If \( \alpha < \beta \), then \( \kappa_\alpha < \kappa_\beta \).

Proof. We will use remark 1024 on \( \beta \).

- The condition \( \alpha < \beta \) is vacuously false for the base case \( \beta = 0 \), hence by (EFQ) the statement vacuously holds.

- Suppose that \( \alpha < \beta \) and \( \kappa_\alpha < \kappa_\beta \). We then have \( \alpha < \text{succ}(\beta) \) and, since, \( \kappa_{\text{succ}(\beta)} > \kappa_\beta \), also \( \kappa_\alpha < \kappa_{\text{succ}(\beta)} \).

- Let \( \lambda \) be a limit ordinal and suppose that the proposition holds for all \( \beta < \lambda \) and for arbitrary \( \alpha \). Then \( \kappa_\beta \subseteq \kappa_\lambda \) for every \( \beta < \lambda \), hence \( \kappa_\beta \leq \kappa_\lambda \) by proposition 1013.

Suppose that \( \alpha < \lambda \).
- If there exists some \( \beta_0 < \lambda \) such that \( \alpha < \beta_0 \), clearly \( \kappa_\alpha < \kappa_{\beta_0} \leq \kappa_\lambda \).
- If \( \alpha > \beta \) for all \( \beta < \lambda \), then \( \alpha \) is an upper bound of the set \( \lambda = \{\beta \mid \beta < \lambda\} \). Hence, \( \alpha \geq \lambda \), which contradicts our choice of \( \alpha \).

Therefore, \( \kappa_\alpha < \kappa_\lambda \).

\[ \square \]

Remark 1056 (Cardinal recursion and induction). Just like we have (bounded and unbounded) transfinite recursion and induction on ordinals, we also have transfinite recursion and induction on cardinals.
We only consider cardinals rather than arbitrary ordinals.

In its structured form presented in remark 1024, rather than considering successor ordinals and limit ordinals, we consider successor cardinals and weak limit cardinals. Thus, recursion and induction on cardinals is formally quite different from the equivalent statements for ordinals. The usage of the two is analogous, however.

See proposition 1057 for how this principles is used.

**Proposition 1057.** For every infinite cardinal \( \kappa \) there exists an ordinal \( \alpha \) such that \( \kappa = \aleph_\alpha \).

**Proof.** We will use remark 1056 (Cardinal recursion and induction) on \( \kappa \).

- The base case \( \kappa = 0 \) vacuously holds because 0 is not an infinite cardinal. The actual base case is \( \kappa = \omega = \aleph_0 \), which holds by definition. This case may not seem formally necessary, however we need to consider it separately from the limit case and calling it the “base case” seems most appropriate.
- If \( \kappa = \aleph_\alpha \) and \( \mu \) is the successor cardinal of \( \kappa \), then by definition \( \kappa = \aleph_{\text{succ}(\alpha)} \).
- Finally, let \( \kappa \) be a limit cardinal and let \( \mu = \aleph_{\alpha_\mu} \) for every infinite cardinal \( \mu < \kappa \). Define

\[
\alpha := \bigcup \{ \alpha_\mu | \mu < \kappa \}.
\]

We have

\[
\kappa \overset{(155)}{=} \bigcup \{ \aleph_{\alpha_\mu} | \mu < \kappa \} \overset{(361)}{\leq} \\
\leq \bigcup \{ \aleph_\beta | \beta < \sup \{ \alpha_\mu | \mu < \kappa \} \} = \\
= \bigcup \{ \aleph_\beta | \beta < \alpha \} \overset{1018}{=} \\
= \aleph_\alpha.
\]

If we suppose that \( \kappa < \aleph_\alpha \), then similarly to proposition 1071 there exists some ordinal \( \beta_0 < \alpha \) such that \( \aleph_{\beta_0} > \aleph_{\alpha_\mu} \) for every \( \mu < \kappa \). In particular, proposition 1055 implies that \( \beta_0 > \alpha_\mu \) for every \( \mu < \kappa \). Thus,

\[
\bigcup_{\alpha} \{ \alpha_\mu | \mu < \kappa \} \leq \beta_0 < \alpha,
\]

which is a contradiction. Therefore, \( \kappa = \aleph_\alpha \).

**Corollary 1058.** The cardinal \( \kappa = \aleph_\alpha \) is a weak limit cardinal if and only if \( \alpha \) is a limit ordinal.

In particular, \( \kappa > 0 \) is a weak limit cardinal if and only if \( \mu < \kappa \) implies that \( \nu < \kappa \), where \( \nu \) is the successor cardinal of \( \mu \).
Proof. Clear from definition 1053.

Definition 1059. We will introduce the notion of countability, which generalizes finiteness.

(a) The smallest infinite cardinal is \( \aleph_0 \). Every set with cardinality \( \aleph_0 \) is called countably infinite. The countably infinite sets are precisely those that can be ordered into a sequence.

(b) A set that is either finite or countably infinite is called at most countable.

(c) Any set that strictly dominates \( \aleph_0 \) is called uncountable. The smallest uncountable cardinal is the successor cardinal \( \aleph_1 \) of \( \aleph_0 \).

(d) The cardinality of \( \text{pow}(\aleph_0) \) has a special name — the cardinality of the continuum. It is sometimes denoted by \( c \). See Conjecture 1061 (Continuum hypothesis) for its relation to \( \aleph_1 \).

See remark 1060 for additional terminology that is potentially more ambiguous.

Remark 1060. Some authors, for example [End77, p. 159], use the shorted term countable, however other authors use “countable” to mean “countably infinite”. The terms denumerable and enumerable are also used for “countably infinite” and “at most countable” respectively. This is done in [Aut20, def. 4.4], for example. These terms are also ambiguous unfortunately.

Conjecture 1061 (Continuum hypothesis). The cardinality of the continuum \( c \) is the first uncountable cardinal \( \aleph_1 \).

Compare this to Conjecture 1088 (Generalized continuum hypothesis).

Remark 1062. Conjecture 1061 (Continuum hypothesis) has been shown by Gödel not to be disprovable in ZFC and by Cohen not to be provable in ZFC.

Proposition 1063. The smallest inductive set \( \omega \) is equinumerous with \( \omega \times \omega \).

Proof. We can give a short proof using theorem 1035 (Cantor-Schröder-Bernstein theorem) using the injective functions \( (n, m) \mapsto 2^n3^m \) in one direction and \( n \mapsto (n, 0) \) in the other direction. Proving the injectivity of \( f \), however, requires theorem 27 (Fundamental theorem of arithmetic), and we prove the latter using machinery from ?? ([UNDEFINED]). We will instead give a direct proof with an explicit construction. We will construct a bijective function from \( \omega \times \omega \) to \( \omega \) — the function visualized in fig. 28.

We begin by defining the diagonal in fig. 28. For each natural number \( k \), define the set of pairs that sum to \( k \):

\[
A_k := \{(n, m) \in \omega \times \omega \mid n + m = k\}.
\]

We can use induction to show that \( \text{card}(A_k) = k + 1 \) numbers. That is, \( A_k \) can “fit” \( k + 1 \) numbers. We can now define the function

\[
d : \omega \to \omega
\]

\[
d(n) := \sum_{k=0}^{n} \text{card}(A_k)
\]
that gives us how many numbers we have already “fit” in the first \( n \) diagonals.

It is clear that the point \((n, m)\) lies in \( A_{n+m} \). We want to know how many numbers we have “fit” in the diagonal prior to that, for which we can use \( d(n + m - 1) \). This leads us to the definition

\[
\begin{align*}
    f : \omega \times \omega &\to \omega \\
    f(n, m) := &\begin{cases} 
        0, & n + m = 0 \\
        d(n + m - 1) + n, & n + m > 0.
    \end{cases}
\end{align*}
\]

We will first show that \( f \) is injective using induction on \( n + m \) (that is, on the diagonals). Suppose that \( f(n_1, m_1) = f(n_2, m_2) \) implies \( n_1 = n_2 \) and \( m_1 = m_2 \) for all pairs with a sum less than \( l \). Let \((n_1, m_1)\) and \((n_2, m_2)\) be two points in \( A_l \) such that \( f(n_1, m_1) = f(n_2, m_2) \).

The cases \( l \leq 1 \) are trivial, so suppose that \( l > 1 \). Then

\[
f(n_1, m_1) = d(n_1 + m_1 - 1) + n_1 = l - 1 + d(n_1 + m_1 - 2) + n_1 = l + f(n_1 - 1, m_1).
\]

We can now apply the inductive hypothesis and obtain that \( n_1 = n_2 \) and \( m_1 = m_2 \). This proves injectivity.

To see that \( f \) is surjective, we will use induction on \( k \in \omega \). The base case is again trivial. Now suppose that \( n + m > 0 \) and

\[
f(n, m) = d(n + m - 1) + n = k.
\]

We have two cases:

- If \( m = 0 \), then

\[
f(0, n + m + 1) = d(n + m) + 0 = d(n + m - 1) + n + m = f(n, m) + 1.
\]

- If \( m > 0 \), then

\[
f(n + 1, m - 1) = d(n + m - 1) + n + 1 = f(n, m) + 1.
\]

In both cases we have shown that \( k + 1 = f(n, m) + 1 \) is in the image of \( f \), which concludes the proof of surjectivity (and hence bijectivity).

Lastly, although it is not necessary for the proof, we can expand the definition of \( d \) to see that \( f \) is actually a polynomial:

\[
f(n, m) = \sum_{k=0}^{n+m-1} (k + 1) + n = \sum_{k=1}^{n+m} k + n = \frac{(n + m)(n + m + 1)}{2} + n.
\]

As an added benefit, this polynomial also handles the case \( n = m = 0 \).

\[
\square
\]

**Corollary 1064**. A finite Cartesian product of at most countable sets is at most countable.
Figure 28: Visualization on an integer coordinate grid of the diagonal sets $A_k$ and of the bijective function defined in proposition 1063.

**Proof.** Let $A_1, \ldots, A_n$ be a finite family of at most countable sets.

Suppose that the inductive hypothesis holds for $n$. Countability ensures that there exist injective functions $g : A_1 \times \cdots \times A_n \times \omega$ and $h : A_{n+1} \to \omega$. Denote by $f$ the bijective function from $\omega$ to $\omega \times \omega$ obtained in proposition 1063 and define

$$F : A_1 \times \cdots \times A_n \times A_{n+1} \to \omega$$

$$F(a_1, \ldots, a_n, a_{n+1}) := f(g(a_1, \ldots, a_n), h(a_{n+1}))$$

By proposition 974, the function $F$ is injective as a superposition of injective functions. Therefore, $\omega$ dominates the product $A_1 \times \cdots \times A_n$, i.e., the product is countable. □

**Proposition 1065.** A countably infinite union of countably infinite sets is countably infinite.

**Proof.** Let $\{A_k\}_{k \in \omega}$ be a countably infinite family of countably infinite sets. Define instead the disjoint family

$$B_k := \{(k, a) \mid a \in A_k\}.$$ 

Denote the union of the former family by $A$ and of the latter family by $B$. Since each $A_k$ is countably infinite, so is $A$. Furthermore, there exists an obvious injective function from $B$ to $A$, thus

$$\aleph_0 \leq \text{card}(A) \leq \text{card}(B).$$ (356)

Define the multi-valued mapping

$$G : \omega \to \text{fun}(\omega, B)$$

$$G(k) := \{g : \omega \to B_k \mid g \text{ is bijective}\}.$$ 

This is a total multi-valued function because we have assumed that $B_k$ is countable for every $k \in \mathcal{K}$. **Theorem 986** (Multi-valued selection existence) gives us a single-valued function $G : \omega \to \text{fun}(\omega, B)$. Since the family $\{B_k\}_{k \in \omega}$ is disjoint, $G$ is injective. We can thus uncurry $G$ to obtain a function $g$ from $\omega \times \omega$ to $B$.

To prove that $g$ is injective, suppose that $g(n_1, m_1) = g(n_2, m_2)$. Then

$$G(n_1)(m_1) = g(n_1, m_1) = g(n_2, m_2) = G(n_2)(m_2).$$

538
Note that \( n_1 \neq n_2 \) would lead to a contradiction because \( \{B_k\}_{k \in \omega} \) is a disjoint family. So \( n_1 = n_2 \) and we obtain \( m_1 = m_2 \) since the function \( G(n_1) : \omega \to B_k \) is injective. Therefore, \( g \) itself is also injective.

It is also surjective because for every \( a \in B \) there exists some \( k \in \omega \) such that
\[
a \in B_k = \text{img}(G(k)).
\]

Denote by \( f \) the bijective function from \( \omega \) to \( \omega \times \omega \) obtained in proposition 1063. Then the function \( f \circ g : \omega \to B \) is bijective by proposition 972. Therefore, the union \( B \) is countably infinite. From (356) it follows that \( A \) is also countably infinite. \( \square \)

**Corollary 1066.** An at most countable union of at most countable sets is at most countable.

**Proof.** Let \( \{A_k\}_{k \in K} \) be an at most countable family of at most countable sets. Denote their union by \( A \). For every \( A_k \) let \( g_k : A_k \to \omega \) be an injective function and define the disjoint union
\[
B_k := A_k \cup \{n \in \omega \mid n \notin \text{img}(g_k)\}
\]
and the bijective function
\[
h_k : B_k \to \omega
\]
\[
h_k(x) := \begin{cases}
(0, g_k(x)), & x \in A_k \\
(1, x), & \text{otherwise}.
\end{cases}
\]

For every \( k \in \omega \setminus K \) instead define
\[
B_k := \{(k, n) \mid n \in \omega\}
\]
and let \( h_k : B_k \to \omega \) be the obvious bijective function.

Then
\[
A = \bigcup_{k \in K} A_k \subseteq \bigcup_{k \in K} B_k \subseteq \bigcup_{k \in \omega} B_k
\]
and the latter is countably infinite by proposition 1065. Therefore, \( A \) is at most countable. \( \square \)
13.8. Transfinite arithmetic

Our purpose is to extend natural number arithmetic to ordinals and cardinals. It turns out that the two are rather different. We will first introduce some additional concepts, however.

**Definition 1067.** Let $A$ and $B$ be sets of ordinals. We say

**Definition 1068.** We recursively define arithmetic operations for arbitrary ordinals as extensions of the corresponding operations of Peano arithmetic.

(a) The sum of $\alpha$ and $\beta$ extends (PA4) and (PA5) with a case for limit ordinals:

\[
\alpha + \beta := \begin{cases} 
\alpha, & \beta = 0 \\
\text{succ}(\alpha + \gamma), & \beta = \text{succ}(\gamma) \\
\sup\{\alpha + \gamma \mid \gamma < \beta\}, & \beta \text{ is a limit ordinal}
\end{cases}
\]  

From proposition 1020 it follows that the in limit case $\alpha + \beta$ is the smallest ordinal strictly larger than $\alpha + \gamma$ for any $\gamma < \beta$.

(b) Analogously, the product of $\alpha$ and $\beta$ extends (PA6) and (PA7):

\[
\alpha \cdot \beta := \begin{cases} 
0, & \beta = 0 \\
\alpha \cdot \gamma + \alpha, & \beta = \text{succ}(\gamma) \\
\sup\{\alpha \cdot \gamma \mid \gamma < \beta\}, & \beta \text{ is a limit ordinal}
\end{cases}
\]  

(c) Exponentiation extends definition 447(e):

\[
\alpha^\beta := \begin{cases} 
1, & \beta = 0 \\
\alpha^\gamma \cdot \alpha, & \beta = \text{succ}(\gamma) \\
\sup\{\alpha^\gamma \mid \gamma < \beta\}, & \beta \text{ is a limit ordinal}
\end{cases}
\]  

**Remark 1069.** For any ordinal $\alpha$ we have

\[
\text{succ}(\alpha) = \text{succ}(\alpha + 0) = \alpha + \text{succ}(0) = \alpha + 1.
\]

We will occasionally use the later notation. Note that for infinite ordinals $\text{succ}(\alpha) = 1 + \alpha$ as discussed in example 1074. This is an extension of remark 9.

**Proposition 1070.** Ordinal addition has the following monotonicity properties:

(a) Left addition is strictly monotone:

\[
\alpha < \beta \text{ implies } \gamma + \alpha < \gamma + \beta.
\]  

540
(b) **Right addition is nonstrictly monotone:**

\[ \alpha < \beta \text{ implies } \alpha + \gamma \leq \beta + \gamma. \]  

(361)

*See example 1074 for examples where the strict inequality fails.*

**Proof.**

**Proof of 1070 (a).** We proceed by induction on \( \beta \).

- The condition \( \alpha < \beta \) is vacuously false for the base case \( \beta = 0 \), hence by (EFQ) the statement vacuously holds.

- Fix some nonzero \( \beta \) and some \( \alpha < \beta \). If \( \gamma + \alpha < \gamma + \beta \), then

\[ \gamma + \alpha < \gamma + \beta < \text{succ}(\gamma + \beta) = \gamma + \text{succ}(\beta). \]

Since \( \beta < \text{succ}(\beta) \), we have used the inductive hypothesis to conclude that

\[ \alpha < \text{succ}(\beta) \text{ implies } \gamma + \alpha < \gamma + \text{succ}(\beta). \]

- Let \( \lambda \) be a limit ordinal and suppose that (360) holds for all \( \beta < \lambda \). Let \( \alpha < \lambda \). Then \( \text{succ}(\alpha) < \lambda \) since \( \lambda \) is a limit ordinal and thus

\[ \gamma + \alpha < \gamma + \text{succ}(\alpha) \leq \text{sup}\{\gamma + \beta \mid \beta < \lambda\} = \gamma + \lambda. \]

**Proof of 1070 (b).** We proceed by induction on \( \gamma \).

- The base case \( \gamma = 0 \) is vacuous.

- If \( \alpha + \gamma < \beta + \gamma \), then

\[ \alpha + \text{succ}(\gamma) \overset{(357)}{=} \text{succ}(\alpha + \gamma) \overset{1014}{<} \text{succ}(\beta + \gamma) \overset{(357)}{=} \beta + \text{succ}(\gamma). \]

- Let \( \lambda \) be a limit ordinal and suppose that the lemma holds for every \( \gamma < \lambda \). That is, for every \( \gamma < \lambda \) we have

\[ \alpha + \gamma < \beta + \gamma. \]

Thus,

\[ \alpha + \lambda = \text{sup}\{\alpha + \gamma \mid \gamma < \lambda\} \leq \text{sup}\{\beta + \gamma \mid \gamma < \lambda\} = \beta + \lambda. \]

We cannot make a stronger conclusion here — see example 1074 for a counterexample.

\[ \square \]

**Proposition 1071.** For any two ordinals \( \alpha \) and \( \beta \) it holds that \( \alpha \leq \beta \) if and only if there exists an ordinal \( \gamma \) such that \( \alpha + \gamma = \beta \). This ordinal is unique and satisfies \( \gamma \leq \beta \).

The strict inequality \( \alpha < \beta \) holds if and only if \( \gamma \neq 0 \).
Proof.

**Proof of sufficiency.** By definition $\beta + 0 = \beta$, hence we are not interested in the case $\alpha = \beta$. That is, we will only consider the case $\alpha < \beta$.

We will first show uniqueness of $\gamma$. Suppose that $\alpha + \gamma_1 = \beta = \alpha + \gamma_2$. From (1070 (a)) it follows that if either $\gamma_1 < \gamma_2$ or $\gamma_1 > \gamma_2$, we would have a strict inequality. Hence, it only remains for $\gamma_1 = \gamma_2$ to hold.

We now use induction on $\beta$ to prove the existence of $\gamma$.

- The condition $\alpha < \beta$ is vacuously false for the base case $\beta = 0$, hence by (EFQ) the statement vacuously holds.

- Suppose that $\alpha < \beta$ and that there exists a unique $\gamma \leq \beta$ such that $\alpha + \gamma = \beta$. Then
  \[ \alpha + \text{succ}(\gamma) = \text{succ}(\alpha + \gamma) = \text{succ}(\beta). \]

Since $\alpha < \beta$ and $\beta < \text{succ}(\beta)$, we have used the inductive hypothesis to conclude that
\[ \alpha < \text{succ}(\beta) \implies \exists \delta < \text{succ}(\gamma). \alpha + \delta = \text{succ}(\beta). \]

Furthermore, since $\gamma \leq \beta$, then also $\text{succ}(\gamma) \leq \text{succ}(\beta)$.

- Suppose that $\lambda$ is a limit ordinal, $\alpha < \lambda$ and for each $\beta < \lambda$ there exists some $\gamma_\beta \leq \beta$ such that $\alpha + \gamma_\beta = \beta$. Define
  \[ \gamma := \sup\{\gamma_\beta | \beta < \lambda\}. \]

By proposition 1020 we have that $\gamma$ is an ordinal and that $\gamma_\beta \leq \gamma$ for every $\beta < \lambda$. Thus,
\[ \lambda = \sup\{\beta | \beta < \lambda\} = \sup\{\alpha + \gamma_\beta | \beta < \lambda\} \leq \sup\{\alpha + \delta | \delta < \text{succ}(\gamma)\} = \alpha + \gamma. \]

Aiming at a contradiction, suppose that the strict inequality holds. That is, suppose that $\lambda < \alpha + \gamma$. Then there exists some $\delta_0 < \gamma$ such that $\alpha + \delta_0 > \alpha + \gamma_\beta$ for any $\beta < \lambda$. It follows from proposition 1228 that $\gamma_\beta < \delta_0$ for any $\beta < \lambda$ and thus
\[ \sup\{\gamma_\beta | \beta < \lambda\} \leq \delta_0 < \gamma. \]

The obtained contradiction shows that such an ordinal $\delta_0$ cannot exist and hence $\lambda = \alpha + \gamma$.

Furthermore, since $\gamma_\beta \leq \beta$ for each $\beta < \lambda$, we have
\[ \gamma = \sup\{\gamma_\beta | \beta < \lambda\} \leq \sup\{\beta | \beta < \lambda\} = \lambda. \]
Proof of necessity. Suppose that $\alpha$, $\beta$ and $\gamma \leq \beta$ are ordinals and that $\alpha + \gamma = \beta$. Obviously $\gamma = 0$ implies that $\alpha = \beta$. If $\gamma > 0$, then from (360) it follows that 
$$\beta = \alpha + \gamma > \alpha + 0 = 0.$$ 

Proposition 1072. Ordinal number addition is associative and left cancellative.
As in proposition 1012, we adapt the corresponding axioms due to theorem 1019 (Burali-Forti paradox). The more concrete result is:

(a) For any three ordinals $\alpha$, $\beta$ and $\gamma$ we have
$$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma).$$

(b) For any three ordinals $\alpha$, $\beta$ and $\gamma$ such that $\gamma + \alpha = \gamma + \beta$, we have $\alpha = \beta$.

See example 1074 for counterexamples to commutativity.
Compare this with proposition 6 and proposition 1079.

Proof.

Proof of 1072 (a). We will use induction on $\gamma$. Proposition 6 already proves the base and successor cases.
Fix some ordinals $\alpha$ and $\beta$. Let $\lambda$ be a limit ordinal and suppose that
$$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$$
holds for all $\gamma < \lambda$. Then
$$(\alpha + \beta) + \lambda \overset{357}{=} \sup\{(\alpha + \beta) + \gamma \mid \gamma < \lambda\} \overset{\text{ind}}{=} \sup\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\} \overset{496}{=} \sup\{\alpha + \delta \mid \delta < \beta + \lambda\} = \alpha + (\beta + \lambda).$$

Proof of 1072 (b). Follows from proposition 1228 and (360).

Proposition 1073. For any two ordinals $\alpha$ and $\beta$, their sum satisfies
$$\alpha + \beta = \text{ord}(\alpha \cup \beta, <),$$
where $<$ is the lexicographic order on the disjoint union $\alpha \cup \beta$.

Proof. We will explicitly build an order isomorphism between $(\alpha + \beta, \in)$ and $(\alpha \cup \beta, <)$. Define
$$f : (\alpha + \beta) \to (\alpha \cup \beta)$$
$$f(\gamma) := \begin{cases} (\gamma, 0), & \gamma < \alpha \\ (\delta, 1), & \exists \delta. (\gamma = \alpha + \delta). \end{cases}$$
From proposition 1071 it follows that the existence of $\delta$ such that $\gamma = \alpha + \delta$ is equivalent to the condition $\gamma \geq \alpha$. Since $\gamma < \alpha + \beta$, we have $\alpha + \delta < \alpha + \beta$ and from proposition 1072 (b) we have $\delta < \beta$. Therefore, $f$ is a total function. Furthermore, it is single-valued because of the uniqueness of $\delta$.

We will first show that $f$ is a strict order homomorphism. Let $\gamma_1 < \gamma_2$. We have the following possibilities:

- If $\gamma_2 < \alpha$, then $f(\gamma_1) = (\gamma_1, 0) < (\gamma_2, 0) = f(\gamma_2)$.
- If $\gamma_1 \geq \alpha$, then $f(\gamma_1) = (\gamma_1, 1) < (\gamma_2, 1) = f(\gamma_2)$.
- If $\gamma_1 < \alpha \leq \gamma_2$, then $f(\gamma_1) = (\gamma_1, 0) < (\gamma_2, 1) = f(\gamma_2)$.

Therefore, $f$ is a strict order homomorphism and from proposition 1242 it follows that $f$ is an order embedding. Due to corollary 1244, in order to show that $f$ is an order isomorphism it only remains to show that it is a surjective function.

Let $(\gamma, k) \in \alpha \sqcup \beta$.

- If $k = 0$, then $f(\gamma) = (\gamma, k)$ since $\gamma \in \alpha$.
- If $k = 1$, then $\gamma \in \beta$ and by (360) we have $\alpha + \gamma < \alpha + \beta$, so $\alpha + \gamma$ is within the domain of $f$. Furthermore, as shown in proposition 1071, if $\alpha + \delta = \alpha + \gamma$, then $\delta = \gamma$. Thus, $f(\alpha + \gamma) = (\gamma, 1)$.

Therefore, $f$ is an order isomorphism between $(\alpha + \beta, \in)$ and $(\alpha \sqcup \beta, \prec)$ and hence

$$\alpha + \beta = \text{ord}(\alpha \sqcup \beta, \prec).$$

Example 1074. The distinction between (501) and (501) is important. A simple example is provided by any limit ordinal $\lambda$, in particular by $\omega$. The example are inconvenient to demonstrate with the recursive definition, however proposition 1073 eases us.

In particular proposition 1073 highlights that adding one ordinal to another, in fact, “appending” a copy of the second to a copy the first.

It is clear that

$$0 + \lambda = \text{ord}(0 \pitchfork \lambda) = \text{ord}(\lambda) = \lambda.$$

That is, we “append” $\lambda$ to an empty well-ordered set only to obtain $\lambda$ again.

This operation seems different from $1 + \lambda$, which “appends” $\lambda$ to a well-ordered singleton set. But this operation only “shifts” $\lambda$ — the function

$$f : \text{ord}(1 \pitchfork \lambda) \rightarrow \text{ord}(0 \pitchfork \lambda)$$

$$f(k, \gamma) := \begin{cases} (0, 0), & k = 0 \\ (0, \gamma + 1), & k = 1. \end{cases}$$

is an order isomorphism and thus

$$1 + \lambda = \text{ord}(1 \pitchfork \lambda) = \text{ord}(\lambda) = \lambda.$$
What the inequality (501) gives us is that

\[ \lambda \leq 1 + \lambda = \lambda. \]

This inequality is, of course, strict when dealing with finite ordinals exclusively, but for limit ordinals its results may be counterintuitive.

What is more interesting is that, as a consequence of (501), we have \( \lambda < \lambda + 1 \). This can be explained as follows. Instead of “appending” an infinite set to a finite one, we append a finite set to an infinite one. This way \( \lambda \) cannot “absorb” 1 like it does in \( 1 + \lambda \).

As a consequence of this example, addition of ordinals is not commutative and also not right-cancellative.

As discussed in the proof of proposition 1079, this is only a restriction of well-orders and not of the resulting sets themselves.

**Proposition 1075.** For any two ordinals \( \alpha \) and \( \beta \), their product satisfies

\[ \alpha \cdot \beta = \text{ord}(\alpha \times \beta, <), \]

where \( < \) is the lexicographic order on the Cartesian product \( \alpha \times \beta \).

**Proof.** We will build an order isomorphism between \( (\alpha \cdot \beta, \in) \) and \( (\alpha \times \beta, <) \) using recursion on \( \beta \).

- Both sets \( \alpha \cdot 0 \) and \( \alpha \cdot 0 \) are empty and the empty function is an order isomorphism.
- Suppose that \( f : \alpha \cdot \beta \to \alpha \times \beta \) is an order isomorphism. We construct the function

\[
\hat{f} : \alpha \cdot (\beta + 1) \to \alpha \times (\beta + 1)
\]

\[
\hat{f} := \begin{cases} 
  f(\gamma), & \gamma < \alpha \cdot \beta \\
  (\delta, \beta), & \exists \delta . (\gamma = \alpha \cdot \beta + \delta).
\end{cases}
\]

In complete analogy with proposition 1071 we can prove that \( \hat{f} \) is an order isomorphism.

- Let \( \lambda \) be a limit ordinal and let \( f_\beta : \alpha \cdot \beta \to \alpha \times \beta \) be an order isomorphism for every \( \beta < \lambda \). Take their union

\[
f := \bigcup \{f_\beta \mid \beta < \lambda \}.
\]

The uniqueness of each \( f_\beta \) from theorem 1031 shows that \( f_\beta_1 \subseteq f_\beta_2 \) for each pair \( \beta_1 < \beta_2 \). Therefore, the union \( f \) is a single-valued partial function. It is also total because every ordinal \( \gamma < \alpha \cdot \lambda \) is contains in the image of the function \( f_{\gamma+1} \).

The function \( f \) is an order embedding by construction. It is also surjective because, if \( (\delta, \beta) \in \alpha \times \lambda \), then from the successor step we can conclude that \( f(\alpha \cdot \beta + \delta) = (\delta, \beta) \).

Therefore, \( f \) is an order isomorphism. 

\[ \square \]
Proposition 1076. Similarly to proposition 1072 for ordinal number addition, multiplication is also associative and left cancellative:

(a) For any three ordinals \( \alpha, \beta \) and \( \gamma \) we have
\[
(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma).
\]

(b) For any three ordinals \( \alpha, \beta \) and \( \gamma \) such that \( \gamma \cdot \alpha = \gamma \cdot \beta \), we have \( \alpha = \beta \).

Compare this with proposition 8 and proposition 1080.

Proof.

Proof of 1076 (a). Associativity follows easily from the obvious isomorphisms between \((\alpha \times \beta) \times \gamma \) and \(\alpha \times (\beta \times \gamma)\).

Proof of 1076 (b). Now suppose that \( \gamma \cdot \alpha = \gamma \cdot \beta \). Let \( f \) the unique order isomorphism between \( \gamma \times \alpha \) and \( \gamma \times \beta \).

Suppose that \( \alpha \neq \beta \). Without loss of generality, suppose that \( \beta \subseteq \alpha \). Then \( f\big|_{\gamma \times \alpha} \) is the identity mapping and hence any set from \( \gamma \times (\beta \setminus \alpha) \) would make \( f \) not injective.

The obtained contradicts shows that \( \alpha = \beta \). \(\square\)

Example 1077. We already know that \( \omega \) is a limit ordinal. It is clear from example 1074 that \( \omega + n \) is a successor ordinal for every natural number \( n \).

What about \( \omega + \omega \)? This corresponds to “placing” two copies of the natural numbers one after another.

Suppose that \( \omega + \omega = \omega \cdot 2 \) is the successor of \( \omega \). Then \( \alpha < \omega + \omega \) and we can show by induction on the natural numbers that \( \alpha + n < \omega + \omega \). But \( \alpha + 1 = \omega \) by assumption, which contradicts trichotomy of ordinals.

Therefore, \( \omega + \omega \) is a limit ordinal. Furthermore, \( \omega + \omega \) is the second smallest limit ordinal since only ordinals of the form \( \omega + n \) for nonzero finite \( n \) satisfy \( \omega < \omega + n < \omega + \omega \).

Both ordinals \( \omega \) and \( \omega + \omega \) are countable by proposition 1063.

Another limit ordinal is \( \omega \cdot \omega = \omega^2 \). It is also countable by corollary 1064. Actually \( \omega^n \) for any natural number \( n \) is countable by the same theorem.

Therefore, any “polynomial” of the form
\[
\alpha_0 \omega^n + \alpha_{n-1} \omega^{n-1} + \cdots + \alpha_1 \omega + \alpha_0
\]
with countable coefficients is also countable.

Definition 1078. We will define arithmetic operations for them. Unlike in definition 1068, we will directly define the operations as cardinal numbers of some sets rather than via some form of recursion.

Fix two ordinals \( \kappa \) and \( \mu \).

(a) Based on proposition 1073, we define their sum as
\[
\kappa + \mu := \text{card}(\kappa \amalg \mu),
\]
where \( \kappa \amalg \mu \) is their disjoint union.
(b) Based on proposition 1075, we define their **product** as
\[ \kappa \cdot \mu := \text{card}(\kappa \times \mu). \]

(c) We define **exponentiation** as
\[ \kappa^\mu := \text{card}(\text{fun}(\kappa, \mu)). \]

**Proposition 1079.** Cardinal number addition is associative, commutative and cancellative.
Compare this with proposition 6 and proposition 1072.

*Proof.* Associativity and left cancellation is inherited from the ordinals. Commutativity and right cancellation hold because we are considering arbitrary bijective functions rather than the more restrictive order isomorphisms. Indeed, \( \kappa \!\!\parallel \!\!\mu \) and \( \mu \!\!\parallel \!\!\kappa \) may have different order types as demonstrated in example 1074, however there is an obvious bijective function between them.

**Proposition 1080.** Cardinal number multiplication is associative, commutative and cancellative.
Compare this with proposition 8 and proposition 1072.

*Proof.* The result follows from the same considerations as in proposition 1079.

**Proposition 1081.** For every cardinal \( \kappa \) we have

\[ (a) \quad \kappa + \kappa = 2\kappa. \]
\[ (b) \quad \kappa \cdot \kappa = \kappa^2. \]

*Proof.*

**Proof of 1081 (a).** Obviously \( \kappa \!\!\parallel \!\!\kappa = 2 \times \kappa. \)

**Proof of 1081 (b).** The function
\[ T : \text{fun}(2, \kappa) \to \kappa \times \kappa \]
\[ T(f) := (f(0), f(1)) \]

is clearly injective. It is also surjective because for any ordered pair \((\gamma, \delta) \in \kappa \times \kappa\) we can define the function
\[ f : 2 \to \kappa \]
\[ f(k) := \begin{cases} \gamma, & k = 0 \\ \delta, & k = 1 \end{cases} \]

Then \( T(f) = (\gamma, \delta). \)

**Lemma 1082.** If \( \kappa \) is an infinite cardinal, then \( \kappa = \kappa^2. \)
Proof. Obviously \( \kappa \leq \kappa^2 \). Suppose that the lemma does not always hold and let \( \kappa \) be the smallest ordinal for which \( \kappa < \kappa^2 \).

Consider the well-ordered set

\[(\kappa \times \kappa \times \kappa, \prec),\]

where \( \prec \) denotes the corresponding lexicographic order. The set is well-ordered as a consequence of proposition 1001.

Define the subset

\[ S := \{ (\alpha, \beta, \gamma) \in \kappa \times \kappa \times \kappa \mid \alpha = \max \{ \beta, \gamma \} \}. \]

The elements of \( S \) are determined exactly by any two of its three coordinates, hence \( S \) is equinumerous with \( \kappa \times \kappa \). Since \( \kappa < \kappa^2 \), there exists some initial segment

\[ S \prec (\alpha_0, \beta_0, \gamma_0) = \{ (\alpha, \beta, \gamma) \in S \mid \alpha < \alpha_0 \text{ and } \beta < \beta_0 \text{ and } \gamma < \gamma_0 \}. \]

Note that \( \alpha_0^2 = \alpha_0 \) since \( \alpha_0 < \kappa \). Then

\[ \kappa = \text{card}(S \prec (\alpha_0, \beta_0, \gamma_0)) \leq \text{card}(S \prec (\alpha_0, \alpha_0, \alpha_0)) = \alpha_0^2 = \alpha_0 < \kappa, \]

which is a contradiction.

Therefore, \( \kappa = \kappa^2 \) for every infinite cardinal \( \kappa \).

**Proposition 1083.** Unlike ordinal arithmetic with its intricacies like example 1074, cardinal arithmetic has a simpler behavior:

(a) If \( \kappa \) and \( \mu \) are finite cardinals, then \( \kappa + \mu \) and \( \kappa \cdot \mu \) are the familiar operations on natural numbers.

(b) If either \( \kappa \) or \( \mu \) is infinite, then

\[ \kappa + \mu = \max \{ \kappa, \mu \}. \]

If, additionally, both are nonzero, then

\[ \kappa \cdot \mu = \max \{ \kappa, \mu \}. \]

**Proof.**

**Proof of 1083 (a).** The addition of ordinals defined in definition 1068 (a) is an extension of addition of natural numbers, hence the two are equivalent for finite ordinals. The equivalence with cardinal addition defined in definition 1078 (a) comes from proposition 1073 and the fact that every finite ordinal is a cardinal as demonstrated in proposition 1039.

Analogously, equivalence of cardinal and ordinal multiplication follows from proposition 1075.
Proof of 1083 (b). Suppose that either $\kappa$ or $\mu$ is infinite and let $\nu := \max\{\kappa, \mu\}$. The cases where either of them is zero are trivial, hence suppose that both are nonzero.

We have

$$\kappa \amalg \mu \subseteq \nu \amalg \nu = 2 \amalg \nu \subseteq \nu \times \nu,$$

hence $\kappa + \mu \leq \nu \cdot \nu = \nu^2$.

Furthermore there exists an obvious injective function from $\nu$ to $\kappa \amalg \mu$ (which is different depending on whether $\nu = \kappa$ or $\nu = \mu$). Therefore,

$$\nu \leq \kappa + \mu \leq \nu^2.$$  

For multiplication we have $\kappa \times \mu \subseteq \nu \times \nu$, hence

$$\nu \leq \kappa \cdot \mu \leq \nu^2.$$  

The rest follows from lemma 1082. \qed

Corollary 1084. The first infinite cardinal $\aleph_0$ is a strong limit cardinal. See also proposition 1099

Proof. Proposition 1083 (a) states that cardinal exponentiation extends natural number exponentiation. Hence, we can conclude that $2^n < \aleph_0$ for any $n < \aleph_0$ since the former are finite and the latter is not. \qed

Lemma 1085. Fix a set $A$. The power set $\text{pow}(A)$ is equinumerous with the set of Boolean-valued functions $\text{pow}(A, \{T, F\})$.

More precisely, then the operator

$$T : \text{fun}(A, \{T, F\}) \to \text{pow}(A)$$

$$T(f) := \{f(x) = T \mid x \in A\}.$$  

is bijective.

Proof. Injectivity is clear. To see surjectivity, fix some subset $B \subset A$ and define

$$f : A \to \{T, F\}$$

$$f(x) := \begin{cases} T, & x \in B \\ F, & \text{otherwise} \end{cases}.$$  

Clearly $f \in \text{fun}(A, \{T, F\})$ and $T(f) = B$. \qed

Proposition 1086. For every set $A$ we have

$$\text{card}(\text{pow}(A)) = 2^{\text{card}(A)}.$$  

549
Proof. The proof is similar to lemma 1085, but much more convoluted.

Let \( \varphi : A \to \text{card}(A) \) be \( \psi : \text{pow}(A) \to \text{card}(\text{pow}(A)) \) bijective functions. Note that
\[
2^{\text{card}(A)} = \text{card}(\text{fun}(\text{card}(A), \{0, 1\}))
\]
by definition. Let \( \theta : \text{fun}(\text{card}(A), \{0, 1\}) \to 2^{\text{card}(A)} \) be a bijective function.

Define the operator
\[
T : 2^{\text{card}(A)} \to \text{card}(\text{pow}(A))
T(p) := \psi(\{\varphi^{-1}(\gamma) \mid \gamma \in \text{card}(A) \text{ and } \theta^{-1}(p)(\gamma) = 1\}).
\]
This operator is bijective since for any \( \delta \in \text{card}(\text{pow}(A)) \) we can define
\[
f : \text{card}(A) \to \{0, 1\}
f(\gamma) := \begin{cases} 1, & \varphi^{-1}(\gamma) \in \psi^{-1}(\delta) \\ 0, & \text{otherwise.} \end{cases}
\]
so that \( T(\theta(f)) = \delta \).

Therefore, \( \text{card}(A) \) and \( \text{card}(\text{pow}(A)) \) are equinumerous. \( \square \)

**Proposition 1087.** If \( \mu \) is the successor cardinal of \( \kappa \), then \( \mu \leq 2^{\kappa} \).

In particular, every strong limit cardinal is a weak limit cardinal.

Furthermore, if \( \kappa \) is infinite, then Conjecture 1088 (Generalized continuum hypothesis) implies that \( \mu = 2^{\kappa} \).

**Proof.** It is clear from theorem 1044 (Cantor’s power set theorem) and proposition 1086 that \( \kappa < 2^{\kappa} \). By definition, \( \mu \) is the smallest cardinal such that \( \kappa < \mu \). Therefore, \( \mu \leq 2^{\kappa} \). \( \square \)

**Conjecture 1088** (Generalized continuum hypothesis). For every ordinal \( \alpha \) we have
\[
\aleph_{\alpha+1} = 2^{\aleph_{\alpha}},
\]
For the definition and the properties of the \( \aleph \) hierarchy, see definition 1053. For the related \( \beth \) hierarchy, see definition 1090.

This is a vast generalization of Conjecture 1061 (Continuum hypothesis) from the case \( \alpha = 0 \) to arbitrary ordinals.

**Corollary 1089.** Conjecture 1088 (Generalized continuum hypothesis) implies that every weak limit cardinal is a strong limit cardinal.

The converse holds in ZFC as shown in proposition 1087.

**Proof.** Follows from proposition 1057 and Conjecture 1088 (Generalized continuum hypothesis). \( \square \)
**Definition 1090.** Similarly to definition 1053, we use transfinite recursion to define, for each ordinal $\alpha$, the cardinal

$$\beth_\alpha : = \begin{cases} \omega, & \alpha = 0 \\ 2^{\beth_\beta}, & \alpha = \text{succ}(\beta) \\ \sup\{\beth_\beta \mid \beta < \alpha\} & \alpha \text{ is a limit ordinal} \end{cases}$$  \hspace{1cm} (362)

Unlike definition 1053, it is able to explicitly build the successor of any member of the hierarchy. **Conjecture 1088** (Generalized continuum hypothesis) states that $\beth_\alpha = \beta_\alpha$ for every ordinal $\alpha$, however in general it is only provable that $\beth_\alpha \leq \beta_\alpha$ — see proposition 1087.
13.9. Von Neumann’s cumulative hierarchy

We will now investigate how we can use set theory itself to build models of set theory. Theorem 1104 contains the important results.

Definition 1091. For every ordinal \( \alpha \) we use remark 1054 (Unbounded transfinite recursion) to define

\[
V_\alpha := \begin{cases} 
\emptyset, & \alpha = 0 \\
\text{pow}(V_\beta), & \alpha = \beta + 1 \\
\bigcup\{V_\beta : \beta < \alpha\}, & \alpha \text{ is a limit ordinal.}
\end{cases}
\] (363)

Each \( V_\alpha \) is called a stage and the index \( \alpha \) of a stage is called its rank. The entire proper class of stages is called the cumulative hierarchy.

If some set \( A \) is a subset of \( V_\alpha \), but not of \( V_\beta \) for any \( \beta < \alpha \), we say that \( \alpha \) is the rank of the set \( A \) and denote it by \( \text{rank}(A) \). We will see in theorem 1093 (Axiom of regularity) that every set has a rank.

Proposition 1092. Without relying on the axiom of foundation and by assuming that ordinals are well-founded by definition, we can prove the following basic properties for Von Neumann’s cumulative hierarchy:

(a) Each stage \( V_\alpha \) is a transitive set.
(b) For any two ordinals \( \alpha < \beta \) we have \( V_\alpha \subseteq V_\beta \).
(c) If \( A \in B \) and both sets have ranks, then \( \text{rank}(A) \subseteq \text{rank}(B) \).
(d) Each stage \( V_\alpha \) is well-founded by set membership.
(e) We have \( \alpha < \beta \) if and only if \( V_\alpha \subseteq V_\beta \).
(f) For every ordinal \( \alpha \) we have \( \text{rank}(\alpha) = \alpha \).
(g) For every stage \( V_\alpha \) we have \( \text{rank}(V_\alpha) = \alpha \).

Proof.

Proof of 1092 (a). The statement is vacuous for \( \alpha = 0 \). Suppose that \( \alpha > 0 \), let \( A \in V_\alpha \) and \( B \in A \). We will show that \( B \in V_\alpha \).

- Suppose that \( \alpha = \beta + 1 \) and that \( V_\beta \) is a transitive set. Then \( A \in V_\alpha \) implies that \( A \subseteq V_\beta \). Thus, \( B \in V_\beta \) and, since \( V_\beta \) is a transitive set, \( B \subseteq V_\beta \).

Therefore, \( B \in V_\alpha = \text{pow}(V_\beta) \).

- Suppose that \( \alpha \) is a limit ordinal and that \( V_\beta \) are transitive sets for every \( \beta < \alpha \). Then \( A \in V_\alpha \) implies that \( A \) belongs to \( V_{\beta_0} \) for some \( \beta_0 < \alpha \). The inductive hypothesis implies that \( A \subseteq V_{\beta_0} \). Therefore, \( A \subseteq V_\alpha \).

Proof of 1092 (b). Let \( \alpha < \beta \) be some ordinals. We will show that \( V_\alpha \subseteq V_\beta \) using induction on \( \beta \).
- Suppose that $\beta$ is a successor ordinal, i.e. $\beta = \mu + 1$ for some $\mu$, and suppose that for every $\alpha < \mu$ we have $V_\alpha \subseteq V_\mu$. Clearly $V_\mu \subseteq V_\beta$ because $V_\mu$ is a subset of itself.

If $\alpha < \mu$, then $V_\alpha \subseteq V_\mu$ by the inductive hypothesis and, since $V_\beta$ is a transitive set by proposition 1092 (a), $V_\alpha \subseteq V_\beta$.

- Suppose that $\beta$ is a limit ordinal. For some fixed $\alpha_0 \in \beta$ we have $V_{\alpha_0} \subseteq V_{\alpha_0+1}$ by what we have already proved. We also have

$$V_\beta = \bigcup_{\alpha < \beta} V_\alpha,$$

hence $V_{\alpha_0} \subseteq V_\beta$.

**Proof of 1092 (c).** Let $A \in B$ be arbitrary sets for which ranks are defined. Trichotomy holds for ordinals, so we have to show that rank$(A) \geq$ rank$(B)$ leads to a contradiction. Denote the ranks by $\alpha$ and $\beta$ for brevity.

We have $A \subseteq V_\alpha$ by definition and $A \subseteq V_\beta$ since $V_\alpha \subseteq V_\beta$ and $V_\beta$ is transitive. If we suppose that rank$(B) <$ rank$(A)$, this would mean that $A$ belongs to a stage below $V_\alpha$, which contradicts the minimality of $\alpha =$ rank$(A)$.

Now suppose that $\beta =$ rank$(B) =$ rank$(A) = \alpha$.

- If $\beta = 0$, then $A \in B$ is impossible.

- If $\beta$ is a successor ordinal of $\alpha$, then $V_\beta =$ pow$(V_\alpha)$. Since $B \subseteq V_\beta$ is a set of subsets of $V_\alpha$, $A \in B$ is a subset of $V_\alpha$. Thus, we have rank$(A) \leq \alpha <$ rank$(A)$, which contradicts the well-foundedness of the ordinal ordering.

- If $\beta$ is a limit ordinal, then $V_\beta = \bigcup\{V_\alpha \mid \alpha < \beta\}$. As a member of $B$, the set $A$ belongs to some lower stage $V_{\alpha_0}$, which again leads to rank$(A) <$ rank$(A)$.

Thus, it remains for rank$(A) <$ rank$(B)$.

**Proof of 1092 (d).** We will use induction on $\alpha$. Suppose that $V_\beta$ is well-founded for every $\beta < \alpha$. Aiming at a contradiction, suppose that there exists an infinitely descending sequence $\{x_k\}_{k=1}^\infty \subseteq V_\alpha$. Then the sequence $\{\text{rank}(x_k)\}_{k=0}^\infty$ of ranks is an infinitely descending set of ordinals. The transitive closure of the underlying set is then an ordinal by proposition 1015. But this contradicts the well-foundedness of ordinals.

Therefore, $V_\alpha$ must be well-founded.

**Proof of 1092 (e).** If $\alpha < \beta$, then from proposition 1092 (b) and proposition 1092 (a) it follows that $V_\alpha \subseteq V_\beta$. We will show that $V_\alpha \not\subseteq V_\beta$. Fix some $\mu < \beta$ and suppose that $V_\alpha \not\subseteq V_\mu$ holds for all $\alpha < \mu$. If $V_\alpha = V_\beta$, this would mean that $V_\alpha \subseteq V_\mu \subseteq V_\beta = V_\alpha$, which is a contradiction. Therefore, $V_\alpha \not\subseteq V_\beta$.

Conversely, suppose that $V_\alpha \not\subseteq V_\beta$. Since trichotomy holds for ordinals and since $V_\alpha \not\subseteq V_\beta$, it is sufficient to show that $\alpha > \beta$ leads to a contradiction. If $\alpha > \beta$, from proposition 1092 (b) it follows that $V_\beta \subseteq V_\alpha$, which implies that $V_\beta \not\subseteq V_\beta$. The obtained contradiction shows that $\alpha < \beta$. 553
**Proof of 1092 (f).** We will use transfinite induction to show that rank(\(\alpha\)) = \(\alpha\) for every ordinal.

- The case \(\alpha = 0\) is trivial because \(\alpha = \emptyset \subseteq \emptyset = V_0\).
- Suppose that \(\alpha = \beta + 1\) is a successor ordinal and rank(\(\beta\)) = \(\beta\). Clearly \(\beta \in \text{pow}(V_\beta) = V_\alpha\) and \(\{\beta\} \in V_\alpha\). Since \(V_\alpha\) is a transitive set, we also have \(\beta \subseteq V_\alpha\). Thus,

  \[
  \alpha = \beta + 1 = \beta \cup \{\beta\} \subseteq V_\alpha.
  \]

- Suppose that \(\alpha\) is a limit ordinal and that for rank(\(\beta\)) = \(\beta\) for all \(\beta < \alpha\). Clearly \(\beta \in V_{\beta+1}\) for any \(\beta < \alpha\), we have

  \[
  \alpha = \{\beta\ \text{is an ordinal} \mid \beta < \alpha\} \subseteq \bigcup \{V_{\beta+1} \mid \beta < \alpha\} = V_\alpha.
  \]

**Proof of 1092 (g).** Clearly \(V_\alpha\) is a subset of itself, hence rank(\(V_\alpha\)) \(\leq \alpha\). In particular, the rank of \(V_\alpha\) exists.

We will use induction on \(\alpha\) to show that rank(\(V_\alpha\)) \(\geq \alpha\).

- For \(\alpha = 0\) this is obvious.
- If rank(\(V_\alpha\)) = \(\alpha\), then,

  \[
  \text{rank}(V_{\alpha+1}) = \text{rank}(\text{pow}(V_\alpha)) \overset{1092 (c)}{>} \text{rank}(V_\alpha) = \alpha.
  \]

- Let \(\lambda\) be a limit ordinal and suppose that rank(\(V_\alpha\)) = \(\alpha\) for any \(\alpha < \lambda\). Then

  \[
  \text{rank}(V_\lambda) \overset{1092 (c)}{\geq} \sup \{\text{rank}(V_\alpha) \mid \alpha < \lambda\} = \sup \{\alpha \mid \alpha < \lambda\} = \lambda.
  \]

Therefore, rank(\(V_\alpha\)) = \(\alpha\). \(\square\)

**Theorem 1093 (Axiom of regularity).** Every set belongs to a stage in von Neumann's cumulative hierarchy.

This statement is called the **axiom of regularity** and in the presence of the other axioms of ZF, it is equivalent to the axiom of foundation. It is much more difficult to state in the language of set theory, however.

**Proof.**

**Proof that axiom of foundation implies axiom of regularity.** Let \(A\) be a set. By proposition 1017, its transitive closure \(\text{cl}^T(A)\) is a transitive set. Define

\[
D := \{B \in \text{cl}^T(A) \mid B \text{ does not belong to the cumulative hierarchy}\}.
\]

We will show that \(D\) is empty. Assume the contrary. Then by the axiom of foundation there exists \(B_0 \in D\) such that \(B_0 \cap D = \emptyset\). Since \(\text{cl}^T(A)\) is a transitive set, \(B_0 \subseteq \text{cl}^T(A)\). Thus, \(B_0\) consists of members \(x\) of \(\text{cl}^T(A)\) that themselves have ranks, i.e. a minimal ordinal \(\beta\) such
that \( x \subseteq V_\beta \). It follows from the axiom schema of replacement that these ordinals form a set. Denote this set by \( C \).

If \( x \in B_0 \), then \( x \subseteq V_\beta \) for some \( \beta \in C \) and \( x \in V_{\beta+1} \). Thus,

\[
B_0 \subseteq \bigcup \{V_{\beta+1} \mid \beta \in C\}.
\]

Denote the union on the right by \( \alpha \). From proposition 1020 it follows that \( \alpha \) is an ordinal strictly larger than the ordinals in \( C \). By proposition 1092 (b) we have that \( V_{\beta+1} \subseteq V_\alpha \) for every \( \beta \in C \). Thus,

\[
B_0 \subseteq \bigcup \{V_{\beta+1} \mid \beta \in C\} \subseteq V_\alpha.
\]

This contradicts our assumption that \( B_0 \) does not belong to the cumulative hierarchy. Therefore, \( D = \emptyset \) and every member of \( cl^T(A) \) also belongs to the cumulative hierarchy. In particular, every member of \( A \) belongs to the cumulative hierarchy.

Define

\[
\mu := \bigcup \{\text{rank}(B) + 1 \mid B \in A\}.
\]

Since \( B \) is a member of the stage with \( \text{rank}(B) + 1 \) for every \( B \in A \), with the same reasoning as above it follows that \( A \subseteq V_\mu \).

**Proof that axiom of regularity implies axiom of foundation.** Let \( A \) be any nonempty set. We will show that there exists a subset of \( A \) that is disjoint from \( A \).

The axiom of regularity ensures that \( A \) belongs to the von Neumann cumulative hierarchy. Let \( B \in A \) be a set with minimal rank.

Suppose that \( B \cap A \) is not empty. Then there exists some set \( C \in A \setminus B \). From proposition 1092 (c) it follows that \( \text{rank}(C) < \text{rank}(B) < \text{rank}(A) \). But \( C \) belongs to \( A \) and has a rank strictly smaller than \( \text{rank}(B) \), which contradicts the minimality of \( \text{rank}(B) \).

The obtained contradiction shows that \( B \cap A = \emptyset \).

**Theorem 1094.** The stage \( V_{\omega+\omega} \) of the von Neumann’s cumulative hierarchy is a standard model of \( Z \), i.e. \( ZFC \) without the axiom schema of replacement.

More generally, a necessary and sufficient condition for \( V_\alpha \) to be a model of \( Z \) is for \( \alpha \) to be a limit ordinal larger than \( \omega \).

**Proof.** The following axioms are automatically satisfied for any stage \( V_\alpha \):

- The validity of the axiom of extensionality is inherited from the metatheory.
- The axiom schema of specification is satisfied because each axiom in the schema defines a subset of \( V_\alpha \) and because \( V_\alpha \) is a transitive set. This does not necessarily require the axiom schema of specification in the metatheory — we only need the subsets of \( V_\alpha \) defined in definition 868.
- The axiom of unions is satisfied because if \( A \in V_\alpha \) and \( C \subseteq \bigcup A \), then there exists some \( B \in A \) such that \( C \subseteq B \subseteq A \). Since \( V_\alpha \) is a transitive set, it follows that \( C \subseteq V_\alpha \).
- The axiom of foundation is satisfied for any \( V_\alpha \) due to proposition 1092 (d). Its validity is also inherited from the metatheory, but we do not actually need the axiom of foundation in the metatheory.
The validity of the **axiom of choice** is inherited from the metatheory.

Indeed, for any family of sets \( \mathcal{A} \subseteq V_\alpha \) there exists a choice function \( c : \mathcal{A} \rightarrow \bigcup \mathcal{A} \). The set \( \{ c(A) \mid A \in \mathcal{A} \} \) then has a lower rank by **proposition 1092 (c)** and thus by **proposition 1092 (e)** it belongs to \( V_\alpha \).

The rest of the axioms are satisfied whenever some easy restrictions are imposed on \( \alpha \):

- **The axiom of power sets** is satisfied by \( V_\lambda \) for a limit ordinal \( \lambda \) if the axiom of power sets holds in the theory.
  
  Indeed, if \( A \subseteq V_\lambda \) and \( A \) has rank \( \beta \), then necessarily \( \beta < \lambda \). Since \( A \subseteq V_\beta \), it follows that \( \text{pow}(A) \subseteq \text{pow}(V_\beta) = V_{\beta+1} \). But \( \beta + 1 < \lambda \) since \( \lambda \) is a limit ordinal. Therefore, \( \text{rank}(\text{pow}(A)) = \beta + 1 < \lambda \). From **proposition 1092 (e)** it follows that \( V_{\beta+1} \subseteq V_\lambda \) and thus \( \text{pow}(A) \subseteq V_\lambda \).

- **The axiom of pairing** is also satisfied by \( V_\lambda \) for a limit ordinal \( \lambda \) if the axiom of pairing holds in the theory.
  
  Let \( A \) and \( B \) be members of \( V_\lambda \). Let \( \beta \) be the larger of their ranks. Then \( A \) and \( B \) are subsets of \( V_\beta \), hence members of \( \text{pow}(V_\beta) = V_{\beta+1} \) and thus the set \( \{ A, B \} \) has rank \( \beta + 2 \).
  
  Since \( \lambda \) is a limit ordinal, we have \( \beta + 2 < \lambda \) and hence \( \{ A, B \} \in V_\lambda \) by **proposition 1092 (e)**.

- **The axiom of infinity** is satisfied by any ordinal \( \alpha > \omega \).
  
  Indeed, by **proposition 1092 (f)** we have \( \omega \subseteq V_\omega \) and by **proposition 1092 (e)** we have \( \omega \in V_\alpha \) for \( \alpha \geq \omega + 1 \).

As discussed in **example 1077**, \( \omega + \omega \) is the smallest ordinal that is both strictly larger than \( \omega \) and is a limit ordinal. Therefore, \( \omega + \omega \) or any larger limit ordinal is a model of ZFC without the axiom of replacement.

*Definition 1095.* The **cofinality** of a preordered set \( (P, \leq) \) is defined as

\[
\text{cf}(P, \leq) := \min\{\text{card}(A) \mid A \text{ is a cofinal subset of } P\}.
\]

*Proposition 1096.* The **cofinality** of an infinite cardinal \( \kappa \) is

\[
\text{cf}(\kappa) = \min\{\text{card}(A) \mid A \text{ is an unbounded subset of } \kappa\}.
\]

*Proof.* Well-foundedness of \( \kappa \) ensures that it is bounded from below, hence a subset \( A \) of \( \kappa \) is bounded from above if and only if it is bounded.

Since \( \kappa \) itself as a limit ordinal by **corollary 1047**, it has no maximum. Hence, it is unbounded.

The above reflections along with **proposition 1247** imply that a subset \( A \) of \( \kappa \) is cofinal if and only if it is unbounded.  

556
Definition 1097. We say that the infinite cardinal $\kappa$ is regular if any of the following equivalent conditions hold:

(a) $\kappa$ it is equal to its own cofinality.

(b) Every unbounded subset of $\kappa$ has cardinality $\kappa$.

Note that the term “regular cardinal” is unrelated to theorem 1093 (Axiom of regularity). If $\kappa$ is not regular, we say that it is singular. Finite ordinals are neither regular nor singular.

Proof.

Proof that 1097 (a) implies 1097 (b). Note that $\text{card}(A) \leq \kappa$ for $A \subseteq \kappa$. Proposition 1096 implies that if $A$ is unbounded, we have $\text{card}(A) \geq \text{cf}(\kappa)$. But $\text{cf}(\kappa) = \kappa$, hence the result follows.

Proof that 1097 (b) implies 1097 (a). If every unbounded subset of $\kappa$ has cardinality $\kappa$, the minimum of all such cardinalities is $\kappa$ and hence $\text{cf}(\kappa) = \kappa$. □

Remark 1098. If $\kappa$ is an uncountable regular (weak or strong) limit cardinal, it is commonly called a (weakly or strongly) inaccessible cardinal.

The assumption of uncountability is sometimes dropped, however. We avoid this ambiguity by being explicit and using “uncountable regular strong limit cardinal” rather than “strongly inaccessible cardinal”.

Furthermore, we often do not need to restrict ourselves to uncountable regular strong limit cardinals. An added benefit to this is that we can utilize the universe of hereditary finite sets $V_\omega$ as a Grothendieck universe in category theory.

Proposition 1099. The first infinite cardinal $\aleph_0$ is a regular.

Proof. The only strict subsets of $\aleph_0$ are either countable or finite and the finite sets are bounded. The only unbounded subsets of $\aleph_0$ have cardinality $\aleph_0$. Hence, $\aleph_0$ equals its own cofinality and is thus regular. □

Lemma 1100. Let $\kappa$ be a regular cardinal and let $R \subseteq \kappa$. If $\text{card}(R) < \kappa$, then $\text{sup} R < \kappa$.

Proof. We shall prove that the set $R$ is bounded from above with respect to the membership ordering of $\kappa$. Indeed, suppose that it is unbounded. Then $\text{card}(R) = \kappa$ since $\kappa$ is regular. But this contradicts our assumption that $\text{card}(R) < \kappa$.

The obtained contradiction shows that $R$ is bounded from above (with respect to membership in $\kappa$). Hence, there exists some ordinal $\rho < \kappa$ that is an upper bound of $R$. The supremum $\text{sup} R$ then satisfies $\text{sup} R \leq \rho < \kappa$. □

Proposition 1101. For any regular cardinal $\kappa$, we have

$$A \subseteq V_\kappa \text{ if and only if } (A \in V_\kappa \text{ and } \text{card}(A) < \kappa).$$

Proof.
Proof of sufficiency. Let $A \subseteq V_\kappa$ and $\text{card}(A) < \kappa$. Define

$$R := \{ r(x) \mid x \in A \}.$$  

Then from lemma 1100 it follows that $\sup R < \kappa$. Denote $\sup R$ by $\rho$. Then

$$x \subseteq V_\rho$$

for every $x \in A$.

Therefore, $A \in V_{\rho+2}$. Since $\kappa$ is an infinite cardinal, from corollary 1047 it follows that it is a limit ordinal and $\rho + 1 < \kappa$; then from proposition 1092 (e) it follows that $A \in V_\kappa$.

Proof of necessity. Follows from proposition 1092 (a).  

Proposition 1102. If $\kappa$ is a regular strong limit cardinal, then $\text{card}(V_\alpha) < \kappa$ for every $\alpha < \kappa$.

Proof. We use theorem 1009 (Bounded transfinite induction) on $\alpha$.

- If $\alpha = 0$, then $V_\alpha = \emptyset$ and hence $\text{card}(V_\alpha) = 0 < \kappa$.
- If $\alpha < \kappa$ and $\text{card}(V_\alpha) < \kappa$, then

$$\text{card}(V_{\alpha+1}) \overset{(363)}{=} \text{card}(\text{pow}(V_\alpha)) \overset{1086}{=} 2^{\text{card}(V_\alpha)}.$$  

Since $\kappa$ is a strong limit, we have $\text{card}(V_{\alpha+1}) < \kappa$.

- Suppose that $\lambda < \kappa$ is a limit ordinal and that $\text{card}(V_\alpha) < \kappa$ for every $\alpha < \lambda$.

From proposition 1092 (e) we have $V_\alpha \subseteq V_{\alpha+1}$ for any $\alpha < \lambda$, hence

$$\text{card}(V_\alpha) \leq \text{card}(V_{\alpha+1}). \quad (364)$$

Define the set

$$C := \{ \text{card}(V_\alpha) \mid \alpha < \lambda \}.$$  

Then

$$\text{card}(V_\lambda) \overset{(363)}{=} \text{card}\left( \bigcup \{ V_\alpha \mid \alpha < \lambda \} \right) \overset{(364)}{=} \sup\{ \text{card}(V_\alpha) \mid \alpha < \lambda \} = \sup C.$$  

From lemma 1100 it follows that $\sup C < \kappa$. Therefore, $\text{card}(V_\lambda) = \sup C < \kappa$.

Corollary 1103. If $\kappa$ is a regular strong limit cardinal, then $\text{card}(A) < \kappa$ for every $A \in V_\kappa$.

Proof. Denote by $\alpha$ the rank of $A$. Then

$$\text{card}(A) \leq \text{card}(V_\alpha) < \kappa,$$

where the last inequality follows from proposition 1102.  

558
**Theorem 1104.** The stage $V_\kappa$ of the von Neumann's cumulative hierarchy is a standard model of ZFC for every uncountable regular strong limit cardinal $\kappa$.

This theorem is an extension of theorem 1094. Thus, assuming that at least one strongly inaccessible cardinal exists, ZFC is consistent since it has a model.

**Proof.** First note that $\kappa$ is necessarily a limit ordinal by corollary 1047. It is also larger than $\omega$ since it is uncountable. Hence, theorem 1094 is satisfied. We must only show that the axiom schema of replacement holds in $V_\kappa$.

Let $A \in V_\kappa$. The axiom schema of replacement requires the image of every definable function from $A$ to $V_\kappa$ to be a member of $V_\kappa$.

Let $\varphi$ be a formula of ZFC not containing $\tau$ nor $\sigma$ as free variables. Suppose additionally that for every $x \in V_\kappa$ there exists a unique $y \in V_\kappa$ such that $\varphi[\xi \to x, \eta \to y] = T$. These conditions ensure that $\varphi$ can be plugged into the axiom schema of replacement. Define the relation

$$f := \{(x, y) \in V_\kappa^2 \mid \varphi[\xi \to x, \eta \to y] = T\}.$$

Then $f$ is a function because of our earlier restrictions on $\varphi$. Since $V_\kappa$ is transitive and $A \in V_\kappa$, it is sufficient to consider the restriction of $f|A$ of $f$. It will not be necessary to do even that since we will explicitly restrict ourselves to values of $f$ on $A$.

Clearly $f[A]$ is a subset of $V_\kappa$ and hence $\text{rank}(f[A]) \leq \kappa$. In order to show that the instance of the axiom schema of replacement with $\varphi$ holds, is sufficient to show that $\text{rank}(f[A]) < \kappa$.

It is clear that

$$\text{card}(f[A]) \leq \text{card}(\text{dom}(f|A)) = \text{card}(A).$$

From corollary 1103 it follows that $\text{card}(A) < \kappa$ and since $\text{card}(f[A]) < \kappa$, from proposition 1101 it follows that $f[A] \in V_\kappa$. □

**Definition 1105.** For this reason, $V_\omega$ is known as the universe of hereditary finite sets.

From corollary 1084 and proposition 1099 it follows that $\omega = \aleph_0$ is a regular strong limit cardinal. Then by corollary 1103, every member of $V_\omega$ is finite. Because $V_\omega$ is a transitive set, every member of every member of $V_\omega$ is also finite. So is every member of every member of every member.

**Proposition 1106.** The universe of hereditary finite sets $V_\omega$ is a standard model of ZFC without the axiom of infinity.

**Proof.** From the proof of theorem 1094 is follows that $V_\omega$ is a model of ZFC without the axiom of infinity and the axiom schema of replacement.

Theorem 1104 shows that being a regular strong limit cardinal is sufficient for $V_\omega$ to satisfy the axiom of replacement. □
13.10. Grothendieck universes

Instead of having one single universe, we can have multiple universes where each is contained in another one. The upside of this is that we can do category theory formally within set theory — see the discussions in definition 1118. The downside of this is that, unlike models of ZFC, models of ZFC+U (ZFC with the axiom that every set is contained in some Grothendieck universe) are much less studied. In particular, this axiom requires the existence of an unbounded hierarchy of regular strong limit cardinal, unlike ZFC for which only one such cardinal is sufficient.

**Definition 1107.** We say that the set $\mathcal{U}$ is a **Grothendieck universe** if it satisfied the following conditions:

- **GU1** It is nonempty.
- **GU2** It is a transitive set.
- **GU3** For any $A \in \mathcal{U}$, the power set $\text{pow}(A)$ also belongs to $\mathcal{U}$.
- **GU4** For any member $A \in \mathcal{U}$ and any $A$-indexed family $\{B_a\}_{a \in A} \subseteq \mathcal{U}$, the union $\bigcup \{B_a \mid a \in A\}$ belongs to $\mathcal{U}$. This is a restriction from unions over completely arbitrary families of sets to those families that can be indexed by members of $A$.

We formalize the entire concept via the following monstrous formula:

$$
\text{IsUniverse}[\mathcal{U}] := \neg \text{IsEmpty}[\mathcal{U}] \land \forall \tau \in \mathcal{U} \left( \text{IsSubset}[\tau, \mathcal{U}] \land \left( \exists \xi \in \mathcal{U} \cdot \text{IsPowerSet}[\xi, \tau] \right) \land \right.
$$

$$
\left. \forall \xi : \forall \eta. \left( \text{IsFun}[\xi, \tau, \mathcal{U}] \land \text{IsImage}[\eta, \xi] \Rightarrow \exists \zeta \in \mathcal{U} \cdot \text{IsUnion}[\zeta, \xi] \right) \right),
$$

where

$$
\text{IsImage}[\rho, \tau] := \forall \xi. \left( \xi \in \rho \leftrightarrow \exists \eta \in \tau. \exists \zeta. \text{IsPair}[\eta, \zeta, \xi] \right). \quad \tau(\zeta) = \xi \text{ for some } \xi \in \text{dom}(\tau)
$$

**Lemma 1108.** Every Grothendieck universe is a superset of universe of hereditary finite sets $V_\omega$.

**Proof.** Let $\mathcal{U}$ be a Grothendieck universe.

- **GU1** ensures that it is nonempty. Then there exists some set $A \in \mathcal{U}$. By **GU3**, $\text{pow}(A) \in \mathcal{U}$. By **GU2**, $\emptyset \in \text{pow}(A) \in \mathcal{U}$ implies $\emptyset \in \mathcal{U}$.

  Finally, from **GU3** by theorem 1009 (Bounded transfinite induction) if follows that $V_{n+1} = \text{pow}(V_n)$ is a member of $\mathcal{U}$ for every $n \in \omega$.

  Therefore, $V_\omega = \bigcup \{V_n \mid n \in \omega\}$ is a subset of $\mathcal{U}$. \qed
**Definition 1109.** The **axiom of universes** states that any set is contained in a Grothendieck universe. Symbolically,

\[
\forall \tau. \exists \nu. \left( \text{IsUniverse}[\nu] \land \tau \in \nu \right).
\] (365)

We usually add this theorem to ZFC and call the resulting logical theory ZFC+U.

**Example 1110.** From theorem 1114 it follows that the universe of hereditary finite sets \( V_\omega \) is a Grothendieck universe.

The existence of other universes cannot be proven in ZFC. For this reason, we use the axiom of universes.

**Proposition 1111.** Suppose that we are working in ZFC+U. Then for any set \( A \), there exists a smallest Grothendieck universe containing \( A \).

More generally, fix a set \( A \). Then there exists a smallest Grothendieck universe containing \( A \).

**Proof.** If no set \( A \) is given, we simply take \( A = \emptyset \) since it must belong to every universe by definition.

We use a trick analogous to proposition 942.

The **axiom of universes** states that there exists at least one universe \( \mathcal{U} \) that contains \( A \). Define

\[ \hat{\mathcal{U}} := \{ x \in \mathcal{U} \mid x \text{ belongs to every Grothendieck universe} \}. \]

Now that we have defined \( \hat{\mathcal{U}} \), it remains to verify that it is itself a universe. To show GU1, note that \( V_\omega \in \mathcal{U} \) by lemma 1108 and hence, by GU2, \( \omega \in \hat{\mathcal{U}} \).

The rest of the verification is trivial. \( \square \)

**Definition 1112.** Suppose \( \mathcal{V} = (V, I) \) is a model of ZFC+U. Let \( \mathcal{U} \) be a fixed Grothendieck universe.

We say that a set \( A \) is \( \mathcal{U} \)-**small** if \( A \in \mathcal{U} \) and \( \mathcal{U} \)-**moderate** if \( A \subseteq \mathcal{U} \). This situation resembles the difference between sets and proper classes described in definition 928.

A set that is not \( \mathcal{U} \)-**small** is called \( \mathcal{U} \)-**large**. Note that any strict superset of \( \mathcal{U} \) is \( \mathcal{U} \)-large, but not \( \mathcal{U} \)-moderate.

Without further context (i.e. in “ordinary mathematics”), we assume that \( \mathcal{U} \) refers to the smallest Grothendieck universe that contains all sets of interest and instead of the terms \( \mathcal{U} \)-large and \( \mathcal{U} \)-small, we simply use the terms **large** and **small**.

In category theory, however, if there is nothing to guarantee the existence of a larger Grothendieck universe, we cannot construct the functor category of \( \mathcal{U} \)-large categories, as discussed in remark 1158. This is the main motivation for the axiom of universes.

**Example 1113.**

- A set is finite if and only if it is \( V_\omega \)-small. A set is \( V_\omega \)-moderate if and only if it is an infinite family of finite sets.

For finite mathematics such as most of combinatorics, we rarely need to work outside of \( V_\omega \).
If $\kappa < \mu$ are regular strong limit cardinals, the stage $V_\kappa$ of von Neumann’s hierarchy is $V_\mu$-small by proposition 1092 (b).

Theorem 1114. The stage $V_\kappa$ of the von Neumann’s cumulative hierarchy is a Grothendieck universe for every regular strong limit cardinal $\kappa$.

Conversely, for every Grothendieck universe $\mathcal{U}$, there exists a regular strong limit cardinal $\kappa$ such that $\mathcal{U} = V_\kappa$.

Proof.

Proof of sufficiency. Let $\kappa$ be a regular strong limit cardinal.

Proof of GU1. Since $\kappa$ is infinite and thus $0 < \kappa$, from proposition 1092 (b) it follows that $V_0 \in V_\kappa$.

Proof of GU2. The set $V_\kappa$ is transitive as shown in proposition 1092 (a).

Proof of GU3. Since $\kappa$ is a limit ordinal, $V_\kappa$ satisfies the axiom of power sets as shown in theorem 1094 and thus if $A \in V_\kappa$, then $\text{pow}(A) \in V_\kappa$.

Proof of GU4. Fix some member $A \in V_\kappa$ and some $A$-indexed family $\{B_a\}_{a \in A} \subseteq V_\kappa$. From corollary 1103 it follows that $\text{card}(A) < \kappa$ and $\text{card}(B_a) < \kappa$ for every $a \in A$. Thus,

$$\text{card}(\{B_a \mid a \in A\}) \leq \text{card}(A) < \kappa$$

and from proposition 1101 it follows that

$$\{B_a \mid a \in A\} \in V_\kappa.$$ 

The union

$$\bigcap \{B_a \mid a \in A\}$$

is then a member of a lower stage, hence it also belongs to $V_\kappa$.

Proof of necessity. Let $\mathcal{U}$ be a Grothendieck universe and let $\alpha$ be the smallest ordinal not in $\mathcal{U}$. We will first show that $\mathcal{U} = V_\alpha$ and gradually prove that $\alpha$ is actually an inaccessible cardinal.

Proof that $V_\beta \in \mathcal{U}$ for ordinals $\beta \in \mathcal{U}$. We will use theorem 1009 (Bounded transfinite induction) on $\beta < \alpha$.

- From lemma 1108 it follows that $\emptyset \in \mathcal{U}$.
- If $\beta < \alpha$ and $V_\beta \in \mathcal{U}$, then by GU3 we have $V_{\beta + 1} = \text{pow}(V_\beta) \in \mathcal{U}$.
  
  We will come back to this step a bit later, but for now note that $V_{\beta + 1} \in \mathcal{U}$ regardless of whether $\beta + 1 \in \mathcal{U}$.
- If $\lambda < \alpha$ is a limit ordinal and $V_\beta \in \mathcal{U}$ for every $\beta < \lambda$, we have

$$V_\lambda \overset{\text{(363)}}{=} \bigcup \{V_\beta \mid \beta < \lambda\},$$

which is a $\lambda$-indexed union of members of $\mathcal{U}$. Since $\lambda \in \mathcal{U}$, by GU3 we have $V_\lambda \in \mathcal{U}$.
Proof that $\alpha$ is a limit ordinal. In the successor case we noted that $V_{\beta+1} \subseteq U$ for every $\beta < \alpha$ regardless of whether $\beta + 1 < \alpha$. Since $\text{rank}(\beta + 1) = \beta + 1$, it follows that $\beta + 1 \in V_{\beta+2} \subseteq U$ and thus by GU2, $\beta + 1 \in U$. Therefore, $\alpha$ cannot be a successor ordinal — if $\alpha = \beta + 1$, then $\beta \in U$ by definition of $\alpha$ and thus $\beta + 1 = \alpha \in U$, which is a contradiction.

Since $\alpha > 0$, it remains for $\alpha$ to be a limit ordinal.

Proof that $V_\alpha \subseteq U$. By GU2, $V_\beta \subseteq U$ for every $\beta < \alpha$. We can conclude that

$$V_\alpha \overset{(163)}{=} \bigcup \{V_\beta \mid \beta < \alpha\} \subseteq U.$$  

In order to show that equality holds, we must first prove that $\alpha$ is a strongly inaccessible cardinal. But this requires some auxiliary results.

Proof that $\{B\} \in U$ for every $B \in U$. By GU3 we have that $\text{pow}(\text{pow}(B)) \in U$. But $\{B\} \subseteq \text{pow}(B)$ and hence $\{B\} \in \text{pow}(\text{pow}(B))$. By GU2, $\{B\} \in U$.

Proof that $\kappa = \alpha$ is a cardinal. Suppose that $\alpha$ is not a cardinal. Indeed, suppose that there exists some $\beta < \alpha$ such that there exists a bijective function $f : \beta \to \alpha$. Then

$$\alpha = \bigcup \{f(\gamma) \mid \gamma < \beta\}$$  

is a $\beta$-indexed union of members of $\alpha$ and hence $\alpha \in \alpha$. But this contradicts corollary 994 (a). Therefore, $\alpha$ is a cardinal. We will henceforth denote it by $\kappa$ to highlight that it is a cardinal.

Proof that $\text{card}(B) \in U$ for every $B \in U$. Let $B \in U$ and let $f : B \to \text{card}(B)$ be a bijective function. Then

$$\text{card}(B) = f[B] = \{f(x) \mid x \in B\} = \bigcup \{f(x) \mid x \in B\}$$  

is a $B$-indexed union of members of $U$ and, by GU4, $\text{card}(B) \in U$.

Proof that $\kappa$ is a strong limit. For every $\beta < \kappa$ by GU3 we have $\text{pow}(\beta) \in U$. We have already shown that $\text{card}(\text{pow}(\beta)) \in U$ and, by proposition 1086, we have

$$\text{card}(\text{pow}(\beta)) = 2^{\text{card}(\beta)} = 2^\beta.$$  

Hence, $2^\beta < \kappa$ and $\kappa$ is a strong limit.

Proof that $\kappa$ is regular. Let $C \subseteq \kappa$ be an unbounded set. We will show that $\text{card}(C) = \kappa$.

Suppose that $\text{card}(C) < \kappa$. Then $\text{card}(C) \in U$ since $\kappa = \alpha$ is the smallest ordinal not contained in $U$. Let $f : \text{card}(C) \to C$ be a bijective function. Then

$$C = \bigcup \{f(\gamma) \mid \gamma < \text{card}(C)\}$$  

and by GU4, $C \in U$.

Since $C$ is unbounded, we have $\sup C \geq \kappa$. But from proposition 1020 is follows that $\bigcup C = \sup C$ and by GU4, $\bigcup C \in U$, which implies that $\sup C < \kappa$.

The obtained contradiction shows that $\text{card}(C) = \kappa$. Since $C$ was an arbitrary unbounded set, it follows that $\kappa$ satisfies definition 1097 (b) and is thus regular.
**Proof that** $V_\kappa = U$. Finally, now that we know that $\kappa$ is a strongly inaccessible cardinal, we can show that equality holds in $V_\kappa \subseteq \mathcal{U}$.

Aiming at a contradiction, suppose that $U \setminus V_\kappa$ is nonempty. By the axiom of foundation, there exists a set $C \in \mathcal{U} \setminus V_\kappa$ such that

$$C \cap (U \setminus V_\kappa) = \emptyset,$$

thus $C \subseteq V_\kappa$. From corollary 1103 it follows that $\text{card}(C) < \kappa$ and from proposition 1101 it follows that $C \in V_\kappa$, which contradicts our choice of $C$ as a member of $U \setminus V_\kappa$.

Therefore, $V_\kappa = U$. 

**Corollary 1115.** Every uncountable Grothendieck universe is a standard model of ZFC.

*Proof.* Follows from theorem 1114 and theorem 1104. 

564
14. Category theory

Category theory studies objects via how they relate to other objects. It shifts the focus from how individual members behave and even has no concept of membership, upon which set theory is based.

This shift is evident from the following diagram, which is actually half of the proof of proposition 1195:

We do still have individual objects, precisely the nodes of the diagram above, however we are only interested in how the nodes are related to each other. Chasing the relations in this diagram individually would require a lot more effort with little gain.

Categories can be defined “from the ground up” so that they may be used without an underlying set theory or logic. For our purposes, it will be more appropriate to define categories via quivers, a.k.a. directed multigraphs. This latter approach will be much more convenient for us, since we are working in ZFC+U and are only interested in categories insomuch as they are helpful to us.

Furthermore, categories are actually the primary motivation for us include the axiom of universes in our metatheory that would otherwise include only the axioms of ZFC. This is discussed further in remark 1135 and remark 1158.
14.1. Categories

**Definition 1116.** A category is a quiver $\mathcal{C}$ equipped with a partial operation $\circ$ on the arrows of $\mathcal{C}$ and another operation $\text{id}$ that selects a distinguished arrow for each vertex.

In tradition regarding forgetful functors, we denote the underlying quiver of $\mathcal{C}$ by $U(\mathcal{C})$.

(a) We call the vertices of the quiver objects and denote the set of all objects by $\text{obj}(\mathcal{C})$.
We will often write $A \in \mathcal{C}$ as a shorthand for $A \in \text{obj}(\mathcal{C})$.

(b) We call the arrows of the quiver morphisms or sometimes maps. If $f$ is a morphism, we call its head its domain $\text{dom}(f)$ and its tail its codomain $\text{codom}(f)$. We denote a morphism from $A$ to $B$ by $f : A \to B$ or $A \rightarrow B$.
We call the set $\mathcal{C}(A,B)$ of all morphisms from $A$ to $B$ a morphism set or hom-set. We use the shorthand $\mathcal{C}(A)$ for $\mathcal{C}(A,A)$. Another established notation is $\text{hom}(A,B)$ instead of $\mathcal{C}(A,B)$.
Both of these notations highlight that $\mathcal{C}(A,B)$, when parameterized by $A$ and $B$, is a functor, as discussed in definition 1165.

(c) We require the composition $\circ$ of the arrows $f$ and $g$ to be defined only if $\text{codom}(f) = \text{dom}(g)$. In this case, we require $g \circ f$ to be a morphism from $\text{dom}(f)$ to $\text{codom}(g)$.

Note how the order of $f$ and $g$ may seem confusing: we write the composition of $f : A \to B$ and $g : B \to C$ as $g \circ f : A \to C$. This is set up so that it matches function composition. The order may seem different compared to multiplication in groups, for example, however definition 1177 shows that this is actually a generalization of multiplication.

This order of composition is used in [Lan94, p. 7], [Lei16, def. 1.1.1] and [Alu09, def. I.3.1].

(d) We denote the identity morphism of an object $A$ by $\text{id}_A$.

The definition of a category additionally requires the following conditions to hold:

C1 For any morphism $f : A \to B$, the identities $\text{id}_A$ and $\text{id}_B$ must satisfy

$$f \circ \text{id}_A = \text{id}_B \circ f = f.$$  \hfill (C1)

C2 Composition must be associative. That is, for each triple of morphism $f : A \to B$, $g : B \to C$ and $h : C \to D$, the following must hold:

$$(h \circ g) \circ f = h \circ (g \circ f).$$  \hfill (C2)

**Example 1117.** Examples of categories include:

- The category $\textbf{Set}$ of small sets and functions defined in definition 1119.
- The category $\textbf{Cat}$ of small categories defined in definition 1143.
• All the categories of small first-order models listed in example 915

• The category \( \textbf{T}_{\text{op}} \) of small topological spaces and continuous functions defined in definition 296.

• For every topological space, the fundamental groupoid defined in definition 295.

• The category \( \textbf{Quiv} \) of small quivers defined in definition 1322.

• For every quiver, the free category defined in definition 1337.

• For every preordered set, the induced category defined in theorem 1186 (Ordered sets as categories).

**Definition 1118.** As can be seen from example 1117, some of the categories we are working with, like \( \textbf{Set} \), contain as objects all small sets. As mentioned in definition 1112, the concept of a small set is defined relative to the smallest Grothendieck universe that suits our needs.

**Theorem 932** (Russell’s paradox) demonstrates that the set of all sets easily leads to a paradox, which is the reason we restrict our attention only to sets within some Grothendieck universe. This universe is implicit by default, however we will occasionally need to make it explicit.

We will say that the category \( \mathcal{C} \) is **locally \( \mathcal{U} \)-small** if the morphism set \( \mathcal{C}(A,B) \) is \( \mathcal{U} \)-small for every pair of objects \( A \) and \( B \). If, in addition, the set \( \text{obj}(\mathcal{C}) \) of objects is also \( \mathcal{U} \)-small, we will say that the category \( \mathcal{C} \) is **\( \mathcal{U} \)-small**. If a category is not \( \mathcal{U} \)-small, we say that it is **\( \mathcal{U} \)-large**.

In particular, finite and **locally finite** categories are ones who are \( V_{\omega} \)-small and \( V_{\omega} \)-locally small for the universe of hereditary finite sets \( V_{\omega} \). This notion of local finiteness is unrelated to local finiteness of graphs defined in definition 1300 (f).

Universes are crucial to be able to do a lot of categorical constructions within set theory, most importantly \( \mathcal{U} \)-large functor categories but also product categories and, as discussed in remark 1135, even the functors themselves.

Note that, even if a category is \( \mathcal{U} \)-small, the category itself as the tuple \((Q,\circ,\text{id})\) from definition 1116 may not be a \( \mathcal{U} \)-small set.

Also note that, in a locally small category, it is possible for the set of all morphisms to be \( \mathcal{U} \)-large. This is impossible for small categories due to GU4.

We sometimes skip the prefix “\( \mathcal{U} \)” if it is unimportant, and simply speak of “large categories” or “locally small categories”.

**Definition 1119.** Suppose that we are given a Grothendieck universe \( \mathcal{U} \), which is safe to assume to be the smallest suitable one as explained in definition 1112.

We denote the category of \( \mathcal{U} \)-small sets by \( \mathcal{U} \cdot \textbf{Set} \) or, if the universe is clear from the context, simply by \( \textbf{Set} \). See definition 1118 for a further discussion of universes and categories.

• The **set of objects** \( \text{obj}(\textbf{Set}) \) is the set of all \( \mathcal{U} \)-small sets, i.e. all members of \( \mathcal{U} \).

• The **set of morphisms** \( \textbf{Set}(A,B) \) from \( A \) to \( B \) is the set \( \text{fun}(A,B) \) of all total single-valued functions from \( A \) to \( B \).
The composition of morphisms is the usual function composition.

The identity morphism on the set $A$ is the identity function

$$\text{id}_A : A \to A$$

$$\text{id}_A(x) := A.$$ 

**Proof of correctness.** To see that $\mathcal{U}\text{-}\text{Set}$ is indeed a category, we verify the conditions $C1$ and $C2$.

**Proof of $C1$.** For every two sets $A, B \in \mathcal{U}$ and every function $f : A \to B$, for all $x \in A$ we have

$$\text{id}_B \circ f(x) = \text{id}_B(f(x)) = f(x) = f(\text{id}_A(x)) = [f \circ \text{id}_A](x).$$

Therefore, $\text{id}_A$ and $\text{id}_B$ satisfy ($C1$).

**Proof of $C2$.** Associativity of function composition is proved in proposition 967 (a).

**Proposition 1120.** We collect here important properties of the category $\mathcal{U}\text{-}\text{Set}$ of $\mathcal{U}$-small sets. Most of them require forward references.

(a) It is a $\mathcal{U}$-large category in the sense of definition 1118 because $\mathcal{U}$ itself is the set of objects and, defined as a quiver with additional operations, the category is a $\mathcal{U}$-large set in the sense of definition 1112.

(b) It is a $\mathcal{U}$-locally small category because $\mathcal{U}$ is a model of ZFC and proposition 985 (p) holds.

(c) All epimorphisms and nonempty monomorphisms split and are precisely the surjective and nonempty injective functions, respectively.

This is stated in proposition 989. See also theorem 1129 (Epimorphisms split in Set).

(d) The empty set $\emptyset$ is an initial object and the singleton set $\{A\}$ is a terminal object for every $A \in \mathcal{U}\text{-}\text{Set}$. No zero objects exist in $\mathcal{U}\text{-}\text{Set}$ by proposition 1133 (d).

This is discussed in example 1131.

(e) The discrete category functor $D : \mathcal{U}\text{-}\text{Set} \to \mathcal{U}\text{-}\text{Cat}$ is left adjoint to the forgetful functor $U : \mathcal{U}\text{-}\text{Cat} \to \mathcal{U}\text{-}\text{Set}$

This is discussed in example 1192 (c).

(f) The products and coproducts are the Cartesian products and the disjoint unions, respectively.

This is stated in proposition 1214.

**Definition 1121.** The opposite category of $\mathcal{C}$ is obtained by “reversing” all arrows. This reversing is merely a relabeling of the domain and codomain — the underlying morphisms are the same. This concept is quite powerful because it allows performing constructions and proofs by duality — see proposition 1124.

Formally, the category $\mathcal{C}^{\text{op}}$ is defined as follows:
The set of objects \( \text{obj}(\mathcal{C}^\text{op}) \) is the set of objects \( \text{obj}(\mathcal{C}) \) of \( \mathcal{C}^\text{op} \).

- The set of morphisms \( \mathcal{C}(A, B) \) is the set \( \mathcal{C}(B, A) \). Thus, any morphism \( f^\text{op} : A \to B \) in the opposite category \( \mathcal{C}^\text{op} \) is a morphism \( f : B \to A \) in \( \mathcal{C}^\text{op} \).

The superscript here is used solely to distinguish between \( f \) being regarded as a morphism of \( \mathcal{C} \) and of \( \mathcal{C}^\text{op} \) — the morphisms in \( \mathcal{C} \) are exactly those of \( \mathcal{C}^\text{op} \), simply relabeled.

- The composition of the morphisms

\[
\begin{align*}
f^\text{op} & \in \mathcal{C}^\text{op}(A, B) = \mathcal{C}(B, A) \\
g^\text{op} & \in \mathcal{C}^\text{op}(B, C) = \mathcal{C}(C, B)
\end{align*}
\]

is the morphism

\[
\begin{align*}
g^\text{op} \circ f^\text{op} & := f \circ g.
\end{align*}
\]

- The identity morphism on the object \( A \in \mathcal{C} \) is again \( \text{id}_A \).

**Remark 1122.** The double-opposite of a category or morphism is obviously the original. This is made precise with the oppositization functor defined in definition 1145.

**Example 1123.** A morphism \( f^\text{op} : A \to B \) in the category \( \text{Set}^\text{op} \) is a function from the set \( B \) to the set \( A \). We cannot apply \( f \) to a point in \( B \) unless \( B \subseteq A \). Thus, we cannot regard, in general, the morphism \( f^\text{op} \) as a function, although only the signature of \( f \) is different from that of \( f^\text{op} \) — their graphs are the same.

**Proposition 1124.** We can extend the principle of duality for preordered sets discussed in definition 1218 (f) to categories. Since we have defined categories in \( \text{ZFC} + U \) rather than as a first-order theory, we will state this principle informally:

If a statement holds for every category, its dual statement obtained by “reversing” all morphisms as in definition 1121, also holds for every category.

See proposition 1128 (b) for how this principle can be utilized easily.

We list here results that heavily utilize this principle. Note that it is now always obvious what exactly needs to reversed in order for this principle to hold. For example, as discussed in definition 1145, for opposite functors we have

\[
[F \circ G]^\text{op} = F^\text{op} \circ G^\text{op},
\]

which is somewhat unexpected.

(a) **Proposition 1127:** A morphism \( f : A \to B \) in \( \mathcal{C} \) is a (split) monomorphism if and only if \( f^\text{op} : B \to A \) in the opposite category \( \mathcal{C}^\text{op} \) is a (split) epimorphism.

In particular, \( f \) is an isomorphism in \( \mathcal{C} \) if and only if \( f^\text{op} \) is an isomorphism in \( \mathcal{C}^\text{op} \).

(b) **Proposition 1132:** An object is initial if and only if it is a terminal object the opposite category.
Proposition 1161: For the opposite of the functor category \([\mathcal{C}, \mathcal{D}]\) we have
\[ \mathcal{C}, \mathcal{D}^{\text{op}} = [\mathcal{C}^{\text{op}}, \mathcal{D}^{\text{op}}]. \]

Proposition 1172: The duals of equivalent categories are equivalent.

Proposition 1190: The functor \(F\) is left adjoint to \(G\) if and only if the dual functor \(F^{\text{op}}\) is right adjoint to \(G^{\text{op}}\).

Proposition 1203: For every cone \((A, \alpha)\) of the diagram \(D\) in \(\mathcal{C}\), \((A, \alpha^{\text{op}})\) is a cocone of \(D^{\text{op}}\) in \(\mathcal{C}^{\text{op}}\).

Even more, for every limit \((L, \pi)\) of \(D\) in \(\mathcal{C}\), \((L, \pi^{\text{op}})\) is a colimit of \(D^{\text{op}}\) in \(\mathcal{C}^{\text{op}}\).

**Definition 1125.** In connection with definition 971 and definition 859, we introduce the following terminology:

(a) The morphism \(g : B \rightarrow C\) is **left-cancellative** if, for any pair of morphisms \(f_1, f_2 : A \rightarrow B\), the equality \(g \circ f_1 = g \circ f_2\) implies \(f_1 = f_2\).

Left-cancellative morphisms are also called **monic morphisms** or **monomorphisms**.

(b) The morphism \(f : A \rightarrow B\) is **left-invertible** if there exists a morphism \(g : B \rightarrow A\) such that \(g \circ f = \text{id}_A\). We call \(g\) a **left inverse** of \(f\).

Using forward references to definition 1148, we can restate this condition by saying that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{f_1} & B \\
\downarrow{f_2} & & \downarrow{g} \\
& C & \\
\end{array}
\]

Left-invertible morphisms are sometimes called **split monomorphisms** because they “split” the identity \(\text{id}_A\) into a composition of \(f\) and \(g\).

(c) Dually, the morphism \(f : A \rightarrow B\) is **right-cancellative** if, for any pair of morphisms \(g_1, g_2 : B \rightarrow C\), the equality \(g_1 \circ f = g_2 \circ f\) implies \(g_1 = g_2\).

Right-cancellative morphisms are also called **epic morphisms** or **epimorphisms**.

(d) The morphism \(g : B \rightarrow A\) is **right-invertible** if there exists a morphism \(f : A \rightarrow B\) such that \(f \circ g = \text{id}_B\). We call \(g\) a **right inverse** of \(f\).

Using forward references to definition 1148, we can restate this condition by saying that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{f} & B \\
\downarrow{\text{id}_B} & & \downarrow{g} \\
& A & \\
\end{array}
\]

Right-invertible morphisms are sometimes called **split epimorphisms** because they “split” the identity \(\text{id}_B\) into a composition of \(g\) and \(f\).
(e) The morphism \( f : A \to B \) is \textbf{fully invertible} if it is both left-invertible and right-invertible. By proposition 1128 (d), in this case, there exists a unique morphism \( f^{-1} : B \to A \) that is a \textbf{two-sided inverse}, i.e. it is both a left inverse and a right inverse.

A fully invertible morphism is usually called an \textbf{isomorphism}. If there exists an isomorphism between \( A \) and \( B \), we say that they are \textbf{isomorphic} and write \( A \cong B \).

(f) A morphism \( f : A \to A \) from an object to itself is called an \textbf{endomorphism}.

(g) A morphism that is both an endomorphism and an isomorphism is called an \textbf{automorphism}.

\textbf{Example 1126.} Proposition 989 characterizes the cancellative and invertible morphisms defined in definition 1125 for \textbf{Set} in terms of injectivity and surjectivity.

A very simple example of a monomorphism which does not split is the empty function with nonempty domain. These are discussed in proposition 989 (a).

\textbf{Theorem 988} (Surjective functions are right-invertible) is important enough to have a categorical interpretation via \textbf{theorem 1129} (Epimorphisms split in \textbf{Set}), where its relation to the \textbf{axiom of choice} is also discussed.

\textbf{Proposition 1127.} A morphism \( f : A \to B \) in \( \mathbf{C} \) is a (split) monomorphism if and only if \( f^{\text{op}} : B \to A \) in the opposite category \( \mathbf{C}^{\text{op}} \) is a (split) epimorphism.

In particular, \( f \) is an \textbf{isomorphism in} \( \mathbf{C} \) if and only if \( f^{\text{op}} \) is an \textbf{isomorphism in} \( \mathbf{C}^{\text{op}} \).

This is part of the duality principles listed in proposition 1124.

\textbf{Proof.} Trivial. \hfill \Box

\textbf{Proposition 1128.} Morphisms have the following basic properties regarding their \textbf{invertibility} (compare to proposition 972):

(a) \textbf{Any left-invertible morphism is left-cancellative.}

In more categorical terms, every split monomorphism is a monomorphism.

(b) \textbf{Any right-invertible morphism is right-cancellative.}

In more categorical terms, every split epimorphism is an epimorphism.

[Let16]
exer. 1.1.13

(c) \textbf{Any morphism has at most one two-sided inverse.}

(d) \textbf{If a morphism is both left-invertible and right-invertible, the two inverses are equal, and the morphism is fully invertible.}

(e) \textbf{The morphism} \( f : A \to B \) \textbf{is a right inverse of} \( g : B \to A \) \textbf{if and only if} \( g \) \textbf{is a left inverse of} \( f \).

(f) \textbf{If a morphism left-cancellative and right-invertible, it is an isomorphism.}

(g) \textbf{If a morphism left-invertible and right-cancellative, it is an isomorphism.}
The composition of two monomorphisms (resp. epimorphisms) is again a monomorphism (resp. epimorphism).

The composition of two split monomorphisms (resp. epimorphisms) is again a split monomorphism (resp. epimorphism).

Proof.

Proof of 1128 (a). Suppose that \( g : B \to C \) is left-invertible with inverse \( h : C \to B \).
Suppose that \( f_1, f_2 : A \to B \) are morphisms such that
\[
g \circ f_1 = g \circ f_2.
\]
Then
\[
f_1 \overset{(c_1)}{=} \text{id}_B \circ f_1 = (h \circ g) \circ f_1 \overset{(c_3)}{=} h \circ (g \circ f_1) = h \circ (g \circ f_2) = \cdots = f_2.
\]

Proof of 1128 (b). This is an exemplar proof using duality. By proposition 1127, every split epimorphism \( f : A \to B \) in \( \mathcal{C} \) is a split monomorphism in \( \mathcal{C}^{\text{op}} \). By proposition 1128 (a), \( f^{\text{op}} \) is a monomorphism. Then gain by proposition 1127, \( f \) is an epimorphism.

Proof of 1128 (c). If \( f : A \to B \) has no inverse, it vacuously has at most one inverse.
Now assume that \( f : A \to B \) has two inverses \( g_1 : B \to A \) and \( g_2 : B \to A \):
\[
g_1 \circ f = \text{id}_A \quad f \circ g_1 = \text{id}_B,
\]
\[
g_2 \circ f = \text{id}_A \quad f \circ g_2 = \text{id}_B.
\]
Then
\[
g_1 \overset{(c_1)}{=} g_1 \circ \text{id}_B = g_1 \circ (f \circ g_2) \overset{(c_2)}{=} (g_1 \circ f) \circ g_2 = \text{id}_A \circ g_2 \overset{(c_1)}{=} g_2.
\]

Proof of 1128 (d). Suppose that \( f : A \to B \) has a left-inverse \( l : B \to A \) and a right-inverse \( r : B \to A \). Then

Proof of 1128 (e). Trivial.

Proof of 1128 (f). Let \( g : B \to A \) be left-cancellative and right-invertible. Let \( f : A \to B \) be a right inverse of \( g \). Then
\[
f = \overset{(c_1)}{=} f \circ \text{id}_A = f \circ (g \circ f) = (f \circ g) \circ f.
\]
Because \( g \) is a left inverse of \( f \), from proposition 1128 (a) it follows that \( f \) is left-cancellative.
Since we have
\[
\text{id}_B \circ f = (f \circ g) \circ f,
\]
it follows that \( f \circ g = \text{id}_B \).
Therefore, \( f \) is a left inverse of \( g \) and hence an isomorphism.

Proof of 1128 (g). The proof is analogous to proposition 1128 (f).
Proof of 1128 (h). Let $g : B \to C$ and $h : C \to D$ be monomorphisms (left-cancellative). Let $f_1, f_2 : A \to B$ be two arbitrary morphisms with codomain $B$. Suppose that

$$(h \circ g) \circ f_1 = (h \circ g) \circ f_2.$$  

Then, by C2,

$$h \circ (g \circ f_1) = h \circ (g \circ f_2).$$

Since $h$ is left-cancellative, it follows that

$$g \circ f_1 = g \circ f_2.$$  

Since $g$ is also left-cancellative, $f_1 = f_2$. Therefore, $h \circ g$ is a monomorphism.

The proof for composition of epimorphisms is identical.

Proof of 1128 (i). Let $f : A \to B$ and $g : B \to C$ be split monomorphisms (left-invertible). Then there exist left inverses $l_f : B \to A$ and $l_g : C \to B$ of $f$ and $g$, respectively. We have

$$(l_f \circ l_g) \circ (g \circ f) \overset{(C2)}{=} l_f \circ (l_g \circ g) \circ f = l_f \circ \text{id}_B \circ f \overset{(C1)}{=} l_f \circ f = \text{id}_A.$$

Therefore, $g \circ f$ is also left-invertible.

The proof for composition of split epimorphisms is identical.

Theorem 1129 (Epimorphisms split in Set). Every epimorphism in $\text{Set}$ splits. That is, all epimorphisms in $\text{Set}$ are split epimorphisms.

Assuming the existence of the Grothendieck universe containing $\text{Set}$, in ZF this theorem is equivalent to the axiom of choice — see theorem 990 (g).

Since not every epimorphism splits in a general category, this theorem is sometimes considered to be a categorical statement of the axiom of choice, which holds in some categories but not in others.

Proof. By proposition 989 (e), a function is an epimorphism if and only if it is surjective. Thus, the theorem is equivalent to theorem 988 (Surjective functions are right-invertible).

Definition 1130. Fix a category $C$.

(a) We call the object $I \in C$ an initial object if for any other object $A \in C$ there exists a unique morphism $f : I \to A$.

(b) Dually, we call the object $T \in C$ a terminal object or final object if for any other object $A \in C$ there exists a unique morphism $f : A \to T$.

The initial and terminal objects are collectively called universal objects.

(c) If $Z$ is both an initial and a terminal object, we say that $Z$ is a zero object. A category with a zero object is called a pointed category.
Example 1131. (a) In the category Set of small sets, for any set $A$ there is a unique empty function from $\emptyset$ to $A$. Therefore, $\emptyset$ is an initial object in Set.

For any set $A$, there is a unique function that contracts any set $B$ to $\{A\}$. Therefore, every singleton set is a final object in Set.

We often denote the initial and terminal objects in Set by $0$ and $1$ respectively, which corresponds to their definition as ordinals.

By proposition 1133 (d), Set has no zero object.

(b) In the category Grp of small groups, the trivial group is a zero object. This holds more generally for pointed sets rather than groups.

Indeed, it can be embedded into any other group and any group can be contracted into the corresponding trivial group. Furthermore, all trivial groups are isomorphic.

Proposition 1132. An object is initial if and only if it is a terminal object the opposite category. This is part of the duality principles listed in proposition 1124.

Proof. Trivial. □

Proposition 1133.

(a) An initial object is unique up to an isomorphism.

(b) Dually, a terminal object is also unique up to an isomorphism.

(c) If a category has an initial and a terminal object and if they are isomorphic, then both are zero objects.

In particular, a zero object is unique up to an isomorphism.

(d) If an initial and a terminal object exists and are not isomorphic, then there exist no zero objects.

Proof.

Proof of 1133 (a). Suppose that $A$ and $B$ are both initial objects in $C$. Then there exist morphisms $f : A \to B$ and $g : B \to A$. Their composition $g \circ f$ is an endomorphism on $A$.

But there exists a unique endomorphism on $A$, which must be the identity $\text{id}_A$. Thus, $g \circ f = \text{id}_A$ and $g$ is a left inverse of $f$.

We can analogously show that $g$ is a right inverse of $f$. Therefore, $f$ is fully invertible, and $A$ and $B$ are isomorphic.

Proof of 1133 (b). If $T'$ and $T''$ are terminal objects in $C$, by proposition 1132, they are initial objects in $C^{\text{op}}$. By proposition 1133 (a), they are isomorphic in $C^{\text{op}}$ and by proposition 1127, they are isomorphic in $C$.

Proof of 1133 (c). Suppose that $A$ is an initial object and that $B$ is a final object in $C$. Let $f : A \to B$ be an isomorphism between them.

Let $C \in C$ be any other object and let $g : C \to B$ be the unique morphism to $B$. Then $f^{-1} \circ g : C \to A$ is a morphism from $C$ to $A$. The inverse $f^{-1} : B \to A$ is unique by...
**Proof of 1133 (d).** By proposition 1133 (c), all zero objects are isomorphic. By proposition 1133 (a), all initial objects are isomorphic and analogously for terminal objects. Hence, if a zero object exists, all initial objects are isomorphic to all terminal objects.

If some initial object is not isomorphic to some terminal object, then by contraposition it follows that no zero object exists. \(\square\)
14.2. Functors

**Definition 1134.** Fix some categories $\mathbf{C}$ and $\mathbf{D}$. A **functor** $F : \mathbf{C} \to \mathbf{D}$ is a quiver homomorphism between the underlying quivers that is compatible with composition and identities. Explicitly, a functor is a family of functions

$$
F_{\text{obj}} : \text{obj}(\mathbf{C}) \to \text{obj}(\mathbf{D})
$$

$$
F_{\text{hom}}(A, B) : \mathbf{C}(A, B) \to \mathbf{D}(F_{\text{obj}}(A), F_{\text{obj}}(B)),
$$

where $F_{\text{hom}}(A, B)$ is a distinct function for every pair of objects $A$ and $B$.

In practice, we usually define the functor as the set

$$
F := F_{\text{obj}} \cup \bigcup \{F_{\text{hom}}(A, B) | A, B \in \text{obj}(\mathbf{C})\}.
$$

Since the domains of all constituent functions are disjoint, $F$ is again a total single-valued function. This allows us to justify the notation $F(A)$ for objects and $F(f)$ for morphisms.

(a) We say that the category $\mathbf{C}$ is the **domain** and $\mathbf{D}$ — the **codomain** of the functor $F$. These are technically not the domain and codomain of $F$ when regarded as a function, however it is consistent with definition 1143.

(b) Similarly to definition 966 (p) for functions, if the domain $\mathbf{C}$ and codomain $\mathbf{D}$ of a functor coincide, we say that it is an **endofunctor**.

The definition of a functor additionally requires the following compatibility conditions to hold:

**CF1** Functors must preserve identities, meaning that for any object $A \in \mathbf{C}$ the following equality must hold:

$$
F(\text{id}_A) = \text{id}_{F(A)}.
$$

**CF2** Functors must preserve composition, meaning that for any pair of morphism $f : A \to B$ and $g : B \to C$ in $\mathbf{C}$,

$$
F(g \circ f) = F(g) \circ F(f).
$$

**Proof of correctness.** The definition (368) ensures that the quiver homomorphism conditions (557) and (558) hold.

Indeed, for any morphism $f : A \to B$ in $\mathbf{C}$ we have

$$
F(\text{dom}(f)) = F(A) = \text{dom}(F(f)),
$$

which implies (557). We also have

$$
F(\text{codom}(f)) = F(B) = \text{codom}(F(f)),
$$

which implies (558).
Remark 1135. It is possible that $\mathcal{C}$ is $\mathcal{U}$-small in the sense of definition 1118, but the functor $F$, as the set (369), is not $\mathcal{U}$-small in the sense of definition 1112. Without using universes, we cannot prove the existence of any functor from the category of smalls sets to itself, for example.

Example 1136. In definition 936, we defined some operations on the category $\textbf{Set}$ of small sets.

(a) The power set $\text{pow} : \textbf{Set} \to \textbf{Set}$ is a canonical example of an endofunctor. Explicitly:

$$\text{pow} : \textbf{Set} \to \textbf{Set},$$

$$\text{pow}(A) := \{S \mid S \subseteq A\},$$

$$\text{pow}(f : A \to B) := (S \mapsto f[S]).$$

We must verify that it is indeed a functor. $\text{CF2}$ is satisfied because

$$\text{pow}(g) \circ \text{pow}(f) = (S \mapsto g[f[S]]) = \text{pow}(g \circ f).$$

The condition $\text{CF2}$ is also obviously satisfied.

The nuance here is that we send every function $f : A \to B$ to its set value $f[S]$ of some subset $S$ of $A$.

(b) The union $\bigcup$ and intersection $\bigcap$ may seem to be good examples of endofunctors in $\textbf{Set}$. Unfortunately, there is no natural way to extend a morphism (function) $f : A \to B$ to a morphism from $\bigcup A$ to $\bigcup B$ or $\bigcap A$ to $\bigcap B$.

Definition 1137. We call the category $\mathcal{D}$ a subcategory of $\mathcal{C}$ if the following hold:

- The underlying quiver $U(\mathcal{D})$ is a subquiver of $U(\mathcal{C})$. That is, every object in $\mathcal{D}$ is an object in $\mathcal{C}$ and every morphism in $\mathcal{D}$ is a morphism in $\mathcal{C}$.

- Composition and identity in $\mathcal{D}$ are restrictions of composition and identity in $\mathcal{C}$.

(a) For every subcategory there exists an inclusion functor $I : \mathcal{D} \to \mathcal{C}$, which sends every object and morphism of $\mathcal{D}$ to itself in $\mathcal{C}$.

(b) We say that $\mathcal{D}$ is a full subcategory if the underlying quiver $U(\mathcal{C})$ is a full subquiver. That is, in case $\mathcal{D}(A,B) = \mathcal{C}(A,B)$ for every pair of objects $A$ and $B$ of $\mathcal{D}$.

By proposition 1152 (c), this is equivalent to the inclusion functor being full.

(c) Every family $\mathcal{D}$ of objects in $\mathcal{C}$ induces a full subcategory $\mathcal{D}$ of $\mathcal{C}$, whose objects are those of $\mathcal{D}$ and whose morphisms are restricted to those whose domain and codomain are both in $\mathcal{D}$.

Remark 1138. We can invert the order of composition in $\text{CF2}$ in the definition of a functor given in definition 1134.
We can replace $\text{CF2}$ with
$$F(g \circ f) = F(f) \circ F(g).$$ (CF2'')

This also requires some other straightforward modifications to the definition of a functor. A functor that satisfies $\text{CF2}'$ rather than $\text{CF2}$ is called contravariant. In this context, a functor satisfying $\text{CF2}$ is called covariant.

Fortunately, a contravariant functor from $\mathcal{C}^{\text{op}}$ to $\mathcal{D}$ is identical to a covariant functor from $\mathcal{C}$ to $\mathcal{D}$. Therefore, there is no formal difference between the two concepts.

The usage of the terms are entirely dictated by context. Unless necessary, we will avoid speaking about contravariant functors to avoid confusion. Some examples where this terminology may be useful are definition 1145, definition 1165 (b) and example 1139.

**Example 1139.** We can try to naïvely define a functor that assigns to a vector space its algebraic dual:

$$F : \text{Vect}_\mathbb{K} \to \text{Vect}_\mathbb{K},$$

$$F(V) := V^*,$$

$$F(f : V \to W) := (\varphi : W \to \mathbb{K} \mapsto \varphi \circ f).$$

Unfortunately, $F(f)$ is supposed to be a morphism from $V^*$ to $W^*$, but is actually a morphism from $W^*$ to $V^*$. This makes $F$ a contravariant functor or, equivalently, a functor from $\text{Vect}_{\mathbb{K}}^{\text{op}}$ to $\text{Vect}_{\mathbb{K}}$.

**Definition 1140.** A discrete category is a category with no morphisms except for the identities. Clearly to any set there corresponds exactly one discrete category and vice versa.

**Example 1141.** Denote by

$$U : \text{Cat} \to \text{Set}$$

the forgetful functor that for any small category $\mathcal{C}$ gives us its set of objects $\text{obj}(\mathcal{C})$. There is also a functor

$$D : \text{Set} \to \text{Cat}$$

that for any small set $A$ gives us the discrete category whose set of objects is $A$.

This is actually an adjunction — see example 1192 (c).

**Proposition 1142.** Functors have the following basic properties:

(a) Functors preserve inverses. For every functor $F : \mathcal{C} \to \mathcal{D}$ and every morphism $f : A \to B$ in $\mathcal{C}$ with a right inverse $g : B \to A$, $F(f)$ is a right inverse of $F(g)$. Similarly, if $g$ is a left inverse of $f$, then $F(g)$ is a left inverse of $F(f)$.

(b) For every functor $F : \mathcal{C} \to \mathcal{D}$ and every isomorphism $f : A \to B$ in $\mathcal{C}$,

$$[F(f)]^{-1} = F(f^{-1}).$$ (370)

(c) Functors preserve isomorphisms. That is, for every functor $F : \mathcal{C} \to \mathcal{D}$, if $f : A \to B$ is an isomorphism in $\mathcal{C}$, $F(f)$ is an isomorphism in $\mathcal{D}$.

Consequently, for every pair of objects $A$ and $B$ in $\mathcal{C}$, from $A \cong B$ it follows that $F(A) \cong F(B)$.

The converse sometimes also holds — see proposition 1152 (h).
Proof.

**Proof of 1142 (a).** Let \( f : A \to B \) be a right inverse of \( g : B \to A \) in \( C \). Then

\[
F(g) \circ F(f) = F(g \circ f) = F(id_A) = \text{id}_{F(A)}.
\]

Thus, \( F(f) \) is a right inverse of \( F(g) \). Since \( g \) is a left inverse of \( f \), automatically \( F(g) \) is a left inverse of \( F(f) \).

**Proof of 1142 (b).** If \( f^{-1} \) is a left inverse of \( f \), by proposition 1142 (a) we have that \( F(f^{-1}) \) is a left inverse of \( F(f) \). But \( F(f^{-1}) \) is also a right inverse, and again by proposition 1142 (a) \( F(g) \) is a right inverse of \( F(f^{-1}) \).

Therefore, \( F(f^{-1}) \) is a two-sided inverse of \( F(f) \). By proposition 1128 (c), it is the only two-sided inverse, hence

\[
[F(f)]^{-1} = F(f^{-1}).
\]

**Proof of 1142 (c).** Follows from proposition 1142 (b).

**Definition 1143.** Suppose that we are given a Grothendieck universe \( \mathcal{U} \), which is safe to assume to be the smallest suitable one as explained in definition 1112.

We denote the category of \( \mathcal{U} \)-small categories by \( \mathcal{U}\text{-}\text{Cat} \) or, if the universe is clear from the context, simply by \( \text{Cat} \). See definition 1118 for a further discussion of universes and categories.

- The set of objects \( \text{obj} (\text{Cat}) \) is the set of all \( \mathcal{U} \)-small categories.
- The set of morphisms \( \text{Cat}(A, B) \) from \( A \) to \( B \) is the set of all functors from \( A \) to \( B \).
- The composition of morphisms is the function composition of the functors regarded as the functions (369). That is, the composition of \( F : C \to D \) and \( G : D \to E \) is the functor

\[
\begin{align*}
[G \circ F] : C &\to E, \\
[G \circ F](A) &:= G(F(A)), \\
[G \circ F](f) &:= G(F(f)).
\end{align*}
\]

- The identity morphism on the category \( C \) is the identity functor

\[
\begin{align*}
id_C : C &\to C, \\
id_C(A) &:= A, \\
id_C(f) &:= f.
\end{align*}
\]

**Proof of correctness.** To see that \( \mathcal{U}\text{-}\text{Cat} \) is indeed a category, we verify the conditions C1 and C2.

**Proof of C1.** For every two \( \mathcal{U} \)-small categories \( C \) and \( D \) and every functor \( F : C \to D \), for every object \( A \in C \) we have

\[
[id_D \circ F](A) = id_D(F(A)) = F(A) = F(id_C(A)) = [F \circ id_C](A)
\]

and analogously for morphisms.

Therefore, \( id_C \) and \( id_D \) satisfy (C1).
Proof of C2. Associativity of functor composition is inherited from the associativity of function composition. □

Definition 1144. For any Grothendieck universe \( \mathcal{U} \), the category of \( \mathcal{U} \)-small categories \( \mathcal{U}-\text{Cat} \) has an initial and a terminal object.

Similarly to how we use the ordinals 0 and 1 to denote the initial and terminal object in the category of sets, we denote the initial category by \( 0 \) and the final category by \( 1 \). Note that the final category is only unique up to an isomorphism. They are identical, however, for all universes \( \mathcal{U} \).

These categories are precisely the discrete categories induced by the ordinals 0 and 1 as described in theorem 1186 (Ordered sets as categories).

[nLa21a] Definition 1145. The opposite functor of \( F : C \to D \) is the functor

\[
F^{\text{op}} : C^{\text{op}} \to D^{\text{op}}
\]

\[
F^{\text{op}}(A) := A
\]

\[
F^{\text{op}}(f^{\text{op}} : B \to A) := [F(f : A \to B)]^{\text{op}}.
\]

For the composition of functors, we then have

\[
[G \circ F]^{\text{op}} = G^{\text{op}} \circ F^{\text{op}}.
\] (373)

This is somewhat in contrast to the general practice of inverting morphisms when taking opposites. Thus, for any Grothendieck universe \( \mathcal{U} \), we have the contravariant oppositization functor

\[
(\cdot)^{\text{op}} : \mathcal{U}-\text{Cat}^{\text{op}} \to \mathcal{U}-\text{Cat}.
\]

As an endofunction on \( \text{obj}(\mathcal{U}-\text{Cat}) \), the oppositization functor is clearly an involution. Dual functors also arise naturally in proposition 1161.

Definition 1146. The image of a functor \( F : C \to D \) is the quiver whose vertex set is

\[
V := \{F(A) \mid A \in C\}
\]

and whose arc set is

\[
A := \{F(f) \mid A, B \in C \text{ and } f \in C(A, B)\}.
\]

This quiver has no categorical structure — it is merely a directed multigraph. As shown in example 1147, imposing a categorical structure naively may fail.

[Wei13] Example 1147.

Consider the functor \( F : C \to D \) from fig. 29.

\[
\bullet \text{ The solid arrows are the morphisms in } C \text{ and their images in } F(C).
\]

\[
\bullet \text{ The dashed arrows denote the action of the functor } F.
\]

580
- The dotted arrow exists in $\mathcal{D}$ as the composition of the other two arrows, however it is missing in the image $F(\mathcal{C})$. Thus, composition is not fully defined in $F(\mathcal{C})$, and $F(\mathcal{C})$ fails to be a category.

**Definition 1148.** Fix a category $\mathcal{I}$, called an index category. A diagram in $\mathcal{C}$ of shape $\mathcal{I}$ is simply a functor $D : \mathcal{I} \to \mathcal{C}$, whose domain is $\mathcal{I}$. We sometimes identify a diagram functor with its image $D(\mathcal{I})$.

It is often convenient to draw graphically the geometric realizations of the quiver $D(\mathcal{I})$. An established convention is to allow multiple vertices representing the same object, which can be achieved formally by actually adjoining new vertices to the quiver and labeling them as in definition 1281. Other established conventions for drawing diagrams include not drawing identity morphisms and adding various visual aids. In this regard, categorical diagrams correspond to the everyday sense of the word “diagram”.

We say that the diagram $D$ over $\mathcal{C}$ commutes if, whenever $p = (f_1, \ldots, f_n)$ and $q = (g_1, \ldots, g_m)$ are two directed paths in $D(\mathcal{I})$ with identical endpoints and either $n > 1$ or $m > 1$, then

$$f_n \circ f_{n-1} \circ \cdots \circ f_2 \circ f_1 = g_m \circ g_{m-1} \circ \cdots \circ g_2 \circ g_1.$$  

We do not really care about how the objects and morphisms in $\mathcal{I}$ are labeled, hence we often use placeholder dots like in (383).

The requirement that the one of the paths is nontrivial, however, is crucial in definition 1215.

**Remark 1149.** Inverting isomorphisms in a commutative diagram may or may not preserve commutativity.
If $p = (f_1, \ldots, f_n)$ and $q = (g_1, \ldots, g_m)$ are two paths in a commutative diagram, and if $f_1$ is invertible, then obviously

$$f_n \circ \cdots \circ f_1 = g_m \circ \cdots \circ g_1$$

if and only if

$$f_n \circ \cdots \circ f_2 = g_m \circ \cdots \circ g_1 \circ f_1$$

and similarly if $f_n$ is invertible.

On the other hand, consider ordinals in $\text{Set}$. Denote by $\iota$ the inclusion maps and by $f : \omega^2 \to \omega$ the bijective map from proposition 1063. Then the following diagram commutes:

![Diagram](374)

but the following does not:

![Diagram](375)

**Definition 1150.** In connection with definition 1125 and definition 971, we introduce the following terminology:

(a) The functor $F : \mathcal{C} \to \mathcal{D}$ is **injective on objects** if the restriction

$$F|_{\text{obj}(\mathcal{C})} : \text{obj}(\mathcal{C}) \to \text{obj}(\mathcal{D})$$

is injective.

That is, for every pair of objects $A$ and $B$ in $\mathcal{C}$, from $F(A) = F(B)$ it follows that $A = B$. If, instead, from $F(A) \cong F(B)$ it follows that $A \cong B$, we say that $F$ is **essentially injective on objects**.

(b) The functor $F : \mathcal{C} \to \mathcal{D}$ is **injective on morphisms** if its restriction to the set

$$\bigcup \{\mathcal{C}(A, B) \mid A, B \in \text{obj}(\mathcal{C})\}$$

of all morphisms is injective.

That is, for every pair of morphisms $f$ and $g$ in $\mathcal{C}$, from $F(f) = F(g)$ it follows that $f = g$. Note that if the morphisms are not parallel, we assume that they are not equal.

(c) The functor $F : \mathcal{C} \to \mathcal{D}$ is **faithful** if it is injective on hom-sets, i.e. for all pairs of objects $A$ and $B$ in $\mathcal{C}$, the restriction of $F$ to $\mathcal{C}(A, B)$ is an injective function.

That is, for every pair of objects $A$ and $B$ in $\mathcal{C}$ and every pair of morphisms $f$ and $g$ in $\mathcal{C}(A, B)$, from $F(f) = F(g)$ it follows that $f = g$.

See proposition 1152 (a) for how faithful functors relate to functors injective on objects or on morphisms.
The functor $F : C \to D$ is **surjective on objects** if the restriction

$$F|_{\text{obj}(C)} : \text{obj}(C) \to \text{obj}(D)$$

is surjective.

That is, for every object $B$ in $D$, there exists at least one object $A$ in $C$ such that $F(A) = B$.

If, instead, there exists at least one object $A \in C$ such that $F(A) \cong B$, we say that $F$ is **essentially surjective on objects**.

Similarly, $F : C \to D$ is **surjective on morphisms** if its restriction to the set of all morphisms is surjective.

That is, for every morphism $g$ in $D$, there exists at least one morphism $f$ in $C$ such that $F(f) = g$.

The functor $F : C \to D$ is **full** if it is surjective on hom-sets, i.e. for all pairs of objects $A$ and $B$ in $C$, the restriction of $F$ to $\text{C}(A, B)$ is a surjective function.

That is, for every pair of objects $A$ and $B$ in $C$ and every morphism $g : F(A) \to F(B)$ in $D$, there exists at least one morphism in $f : A \to B$ in $C$ such that $F(f) = g$.

Finally, $F : C \to D$ is **fully faithful** if it is both full and faithful.

**Proposition 1151.** Functors preserve commutative diagrams and faithful functors also reflect commutative diagrams.

More precisely, let $C$ be an arbitrary category, let $D$ be a diagram in $C$, and let $p = (A, f_1, \ldots, f_n)$ and $q = (A, g_1, \ldots, g_m)$ be two directed paths with the same endpoints in $D$.

For any functor $F : C \to D$, if

$$f_n \circ \cdots \circ f_1 = g_m \circ \cdots \circ g_1,$$

then

$$F(f_n) \circ \cdots \circ F(f_1) = F(g_m) \circ \cdots \circ F(g_1),$$

Conversely, if $F$ is faithful, then (377) implies (376).

**Proof.** Functors preserve composition by $\text{CF}_2$, hence (377) follows from (376) directly.

Now suppose that (377) holds for a faithful functor $F$. $\text{CF}_2$ allows us to reduce (377) to

$$F(f_n \circ \cdots \circ f_1) = F(g_m \circ \cdots \circ g_1).$$

Then, by injectivity of $F$ on the morphism set $\text{C}(\text{dom}(f_1), \text{codom}(f_1))$, (376) holds.

**Proposition 1152.** Functors have the following basic properties regarding their invertibility:

(a) A functor is **injective on morphisms** if and only if it is both **injective on objects** and **faithful**.

(b) A functor is **surjective on morphisms** if and only if it is both **surjective on objects** and **full**.
(c) A subcategory $\mathcal{D}$ of $\mathcal{C}$ is full in the sense of definition 1137 if and only if the inclusion functor $I : \mathcal{D} \rightarrow \mathcal{C}$ is full in the sense of definition 1150 (f).

(d) Any functor preserves left, right inverses and two-sided inverses.

(e) **Faithful** functors reflect composition. That is, for every functor $F : \mathcal{C} \rightarrow \mathcal{D}$, if the following diagram commutes:

$$
\begin{array}{c}
F(A) \\ F(h) \downarrow \\
F(C)
\end{array} \quad \xRightarrow{F(g)} \quad \begin{array}{c}
F(B) \\ F(f) \downarrow \\
F(C)
\end{array}$$

then the following diagram are identities:

$$
\begin{array}{c}
A \\ h \downarrow \\
C
\end{array} \quad \xRightarrow{f} \quad \begin{array}{c}
B \\ g \downarrow \\
C
\end{array}$$

(f) A **faithful** functor reflects monomorphisms and epimorphisms. That is, for every functor $F : \mathcal{C} \rightarrow \mathcal{D}$ and morphism $f : A \rightarrow B$ in $\mathcal{C}$, if $F(f)$ is a monomorphism (resp. epimorphism), so is $f$.

(g) A **fully faithful** functor reflects identities. That is, for every functor $F : \mathcal{C} \rightarrow \mathcal{D}$ and endomorphism $f : A \rightarrow A$ in $\mathcal{C}$, if $F(f) = \text{id}_{F(A)}$, then $f = \text{id}_A$.

(h) A **fully faithful** functor reflects split monomorphisms and split epimorphisms.

That is, for every functor $F : \mathcal{C} \rightarrow \mathcal{D}$ and morphism $f : A \rightarrow B$ in $\mathcal{C}$, if $F(f)$ is a split monomorphism (resp. split epimorphism or isomorphism), so is $f$.

(i) A functor between $\mathcal{U}$-small categories that is both injective and surjective on morphisms is itself an isomorphism in $\mathcal{U}$-\textbf{Cat}.

**Proof.**

**Proof of 1152 (a).**

**Proof of sufficiency.** Let $F : \mathcal{C} \rightarrow \mathcal{D}$ be injective on morphisms. It is trivially faithful since faithfulness is a more restrictive condition.

To see that $F$ is injective on objects, let $A, B \in \mathcal{C}$ and suppose that $F(A) = F(B)$. Then $\text{id}_{F(A)} = \text{id}_{F(B)}$ and $F(\text{id}_A) = F(\text{id}_B) = \text{id}_{F(A)} = \text{id}_{F(B)} = F(\text{id}_B)$.

Since $F$ is injective on morphisms, it follows that $\text{id}_A = \text{id}_B$, hence $A = B$. Thus, $F$ is injective on objects.
**Proof of necessity.** Let $F : C \to D$ be faithful and injective on objects. Let $f : A \to B$ and $g : C \to D$ be morphisms in $C$ such that $F(f) = F(g)$.

Then both $F(f)$ and $F(g)$ have the same domain $F(A) = F(C)$ and codomain $F(B) = F(D)$. Hence, since $F$ is injective on objects, we have $A = C$ and $B = D$.

Thus, $f$ and $g$ are both morphisms from $A$ to $B$. Since $F$ is also faithful, from $F(f) = F(g)$ it follows that $f = g$.

Therefore, $F$ is injective on morphisms.

**Proof of 1152 (b).**

**Proof of necessity.** Let $F : C \to D$ be surjective on morphisms. It is trivially full since fullness is a more restrictive condition.

To see that $F$ is surjective on objects, let $C \in D$. Then there exists some morphism $f : A \to B$ in $C$ such that $F(f) = \text{id}_2$. We thus necessarily have $F(A) = C$ and $F(B) = C$.

**Proof of sufficiency.** Let $F : C \to D$ be full and injective on objects. Let $g : C \to D$ be a morphism in $D$.

Since $F$ is surjective on objects, there exists preimages $A$ of $C$ and $B$ of $D$ under $F$. Thus, $g \in D(F(A), F(B))$.

Since $F$ is also full, there exists some morphism $f : A \to B$ such that $F(f) = g$.

Therefore, $F$ is surjective on morphisms.

**Proof of 1152 (c).** Trivial.

**Proof of 1152 (d).** Let $g : B \to A$ be a left inverse of $f : A \to B$. Then

$$F(g) \circ F(f) \overset{(\text{CF2})}{=} F(g \circ f) = F(\text{id}_A) \overset{(\text{CF})}{=} \text{id}_{F(A)}.$$  

The case of right inverses is similar.

**Proof of 1152 (e).** Suppose that (378) commutes. Then, since $F$ is faithful and thus injective on the morphism set $C(A, C)$, the equality $F(g \circ f) = F(g) \circ F(f) = F(h)$ implies that $g \circ f = h$. Hence, (379) also commutes.

**Proof of 1152 (f).** Let $F(g)$ be a monomorphism and let $f_1, f_2 : A \to B$ be parallel morphisms such that

$$g \circ f_1 = g \circ f_2.$$  

Then, since $F(g)$ is a monomorphism, we have that $F(f_1) = F(f_2)$. Since $F$ is faithful, the restriction $F|_{C(A,B)}$ is injective, and $f_1 = f_2$.

The proof when $F(g)$ is an epimorphism is analogous.

**Proof of 1152 (g).** If $F : C \to D$ is fully faithful, for every object $A$ in $C$, the identity morphism $\text{id}_{F(A)}$ has a unique preimage under $F$. By CF1, this preimage can only be $\text{id}_A$.

**Proof of 1152 (h).** Let $q$ be a left inverse of $F(f)$. Since $F$ is fully faithful, there exists a unique morphism $g : B \to A$ such that $F(g) = q$.

Since

$$F(g) \circ F(f) = \text{id}_{F(A)},$$

by proposition 1152 (g) we have

$$g \circ f = \text{id}_A.$$
Therefore, $g$ is a left inverse of $F(f)$.
The proof for right inverses follows from proposition 1128 (e).
From proposition 1128 (d) it follows that if $F(f)$ is an isomorphism, so is $f$.

**Proof of 1152 (i).** If $F$ is both injective and surjective on morphisms, it is also injective and surjective on objects and hence, as a function, is bijective. Therefore, it is both left and right invertible as a consequence of proposition 989 (f).

**Example 1153.**

(a) The power set functor described in example 1136 is clearly injective on morphisms, hence by proposition 1152 (a), it is also injective on objects and faithful.

It is not full, nor surjective on objects.

(b) The forgetful functor $D : \text{u-Cat} \to \text{u-Set}$ discussed in definition 1140 is surjective on morphisms, hence by proposition 1152 (b), it is also surjective on objects and full.

It is not faithful, nor injective on objects.

**Definition 1154.** Let $F$ and $G$ be parallel functors from the category $C$ to $D$.

A natural transformation $\alpha$ from $F$ to $G$ is an indexed family of

$$\{\alpha_A : F(A) \to G(A)\}_{A \in C} \quad (380)$$

of morphisms in $D$ such that, for every morphism $f : A \to B$ in $C$, the following diagram commutes:

$$\begin{array}{ccc}
F(A) & \xrightarrow{F(f)} & F(B) \\
\downarrow{\alpha_A} & & \downarrow{\alpha_B} \\
G(A) & \xrightarrow{G(f)} & G(B)
\end{array} \quad (381)$$

The morphisms $\alpha_A$ are called the components of $\alpha$. We denote natural transformations by $\alpha : F \Rightarrow G$ and, when used in diagrams, by

$$\begin{array}{ccc}
& & \uparrow{\alpha} \\
C & \xrightarrow{F} & D \\
& & \downarrow{G}
\end{array} \quad (382)$$

**Example 1155.** In definition 1320, we have defined a quiver as a set $V$ of vertices, a set $A$ of arcs and two functions — the head $h : A \to V$ and tail $t : A \to V$ of an arc.

Now consider the following index category $I$:

$$\bullet \xrightarrow{\alpha} \bullet \quad (383)$$

For the sake of readability, we will give the following explicit labels in this category:

$$V \xrightarrow{h} A \quad (384)$$

586
A quiver can then be defined as a functor \( Q : I \to \mathcal{U}\text{-Set} \) to the category \( \mathcal{U}\text{-Set} \) of \( \mathcal{U} \)-small sets (for a fixed Grothendieck universe \( \mathcal{U} \)).

A natural transformation from the quiver \( Q : I \to \mathcal{U}\text{-Set} \) to \( R : I \to \mathcal{U}\text{-Set} \) is then a pair of functions \( f_V : Q(V) \to R(V) \) and \( f_A : Q(A) \to R(A) \) such that the following diagrams commute:

\[
\begin{array}{c}
Q(A) \xrightarrow{Q(h)} Q(V) \\
\downarrow f_A \quad \downarrow f_V \\
R(A) \xrightarrow{R(h)} R(V)
\end{array}
\]

(385)

See example 1159 for how these functors relate to quivers as defined in definition 1320.

Remark 1156. Let \( \mathcal{C} \) be an arbitrary \( \mathcal{U} \)-small category. A natural transformation \( \alpha \) from \( F : \mathcal{C} \to \mathcal{U}\text{-Set} \) to \( G : \mathcal{C} \to \mathcal{U}\text{-Set} \) is then a family of functions \( \{ \alpha_A : F(A) \to G(A) \}_{A \in \mathcal{C}} \).

Suppose that for every two objects \( A \) and \( B \) in \( \mathcal{C} \), the functions \( \alpha_A \) and \( \alpha_B \) agree on \( F(A) \cap F(B) \). This is automatically satisfied in \( F(A) \) and \( F(B) \) are disjoint whenever \( A \neq B \).

We can then take the set-theoretic union of \( \alpha \) to obtain the function

\[
\bigcup_{A \in \mathcal{C}} \alpha_A : \bigcup \{ F(A) \mid A \in \mathcal{C} \} \to \bigcup \{ G(A) \mid A \in \mathcal{C} \}.
\]

Both the domain and codomain are sets as a consequence of GU4, therefore the function is well-defined in the universe \( \mathcal{U} \). Denote it on \( A \) for brevity.

An advantage of this is that we can define a natural transformation to be a function on a general enough set and then prove that its restrictions satisfy (381).

For example, consider the power set functor \( \text{pow} : \mathcal{U}\text{-Set} \to \mathcal{U}\text{-Set} \) discussed in example 1136. The identity function \( \text{id}_{\mathcal{U}\text{-Set}} \) is then a natural transformation from the identity functor \( \text{id}_{\mathcal{U}\text{-Set}} \) to \( \text{pow} \).

Another natural transformation between the same functors is the singleton set operation \( \Sigma \) on sets defined as \( A \mapsto \{ A \} \). Note that, in this context, \( \Sigma \) operates not on the sets \( \text{id}_{\mathcal{U}\text{-Set}}(A) \) and \( \text{pow}(A) \), but on their members. The diagram (381) becomes

\[
\begin{array}{c}
A \xrightarrow{f} B \\
\downarrow \Sigma \\
\text{pow}(A) \xrightarrow{\text{pow}(f)} \text{pow}(B)
\end{array}
\]

(386)

This diagram commutes because, for every function \( f : A \to B \) and every \( x \in A \), we have

\[
f([x]) = \{ f(x) \}.
\]

Definition 1157. Let \( \mathcal{C} \) and \( \mathcal{D} \) be arbitrary categories. The functor category \( [\mathcal{C}, \mathcal{D}] \), also denoted as \( \mathcal{D}^\mathcal{C} \), is defined as follows:
• The set of objects \( \text{obj}([\mathbf{C}, \mathbf{D}]) \) is the set of all functors from \( \mathbf{C} \) to \( \mathbf{D} \).

• The set of morphisms \( [\mathbf{C}, \mathbf{D}](F, G) \) from \( F \) to \( G \) is the set of all natural transformations from \( F \) to \( G \).

• The composition of the morphisms \( \alpha : F \Rightarrow G \) and \( \beta : G \Rightarrow H \) is the natural transformation \( \beta \circ \alpha : F \Rightarrow H \) defined in terms of componentwise morphism composition, i.e.
  \[
  (\beta \circ \alpha)_A := \beta_A \circ \alpha_A.
  \] (387)

• The identity morphism on the functor \( F : \mathbf{C} \rightarrow \mathbf{D} \) is the identity natural transformation \( \text{id}_F : F \Rightarrow F \) with components
  \[
  (\text{id}_F)_A := \text{id}_{F(A)}
  \] (388)

**Proof of correctness.** Just to verify that the composition \( \beta \circ \alpha \) defined in (387) is indeed a natural transformation from \( F \) to \( H \), note that the following diagram trivially commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{F(f)} & F(B) \\
\alpha_A & & \beta_B \\
G(A) & \xrightarrow{G(f)} & G(B) \\
\beta_A & & \beta_B
\end{array}
\] (389)

Now, to see that \( [\mathbf{C}, \mathbf{D}] \) is indeed a category, we verify the conditions C1 and C2, which are in turn inherited from the same conditions on the categories \( \mathbf{C} \) and \( \mathbf{D} \).

**Proof of C1.** For every two functors \( F, G : \mathbf{C} \rightarrow \mathbf{D} \) and natural transformation \( \alpha : F \Rightarrow G \), for every object \( A \in \mathbf{C} \) we have

\[
\text{id}_{G(A)} \circ \alpha_A \overset{(\text{C1})}{=} \alpha_A \overset{(\text{C1})}{=} \alpha_A \circ \text{id}_{F(A)}
\]

Therefore,

\[
\text{id}_G \circ \alpha = \alpha = \alpha \circ \text{id}_F
\]

and, after generalizing, we obtain that (C1) holds in \([\mathbf{C}, \mathbf{D}]\).

**Proof of C2.** For any quadruple \( F, G, H \) and \( T \) of functors from \( \mathbf{C} \) to \( \mathbf{D} \) and every combination of natural transformations \( \alpha : F \Rightarrow G, \beta : G \Rightarrow H \) and \( \gamma : H \Rightarrow T \), for every object \( A \in \mathbf{C} \) we have

\[
(\gamma_A \circ \beta_A) \circ \alpha_A \overset{(\text{C2})}{=} \gamma_A \circ (\beta_A \circ \alpha_A).
\]

Therefore, after generalizing, we obtain that (C2) holds in \([\mathbf{C}, \mathbf{D}]\). \qed
Remark 1158. If \( \mathcal{C} \) and \( \mathcal{D} \) are \( \mathcal{U} \)-large categories in the sense of definition 1118, we cannot construct the functor category \( [\mathcal{C}, \mathcal{D}] \). This is the main motivation for the axiom of universes, which is discussed in definition 1112 and, in relation to category theory, in definition 1118.

Example 1159. In example 1155, we defined quivers as functors from a certain index category \( I \) to \( \mathcal{U} \- \text{Set} \) (for a fixed Grothendieck universe \( \mathcal{U} \)).

There is then an obvious correspondence between quivers as objects of \( \mathcal{U} \- \text{Quiv} \), defined in definition 1320, and quivers as objects in the functor category \( [I, \mathcal{U} \- \text{Set}] \), defined in example 1155. Indeed, given any functor \( Q : I \to \mathcal{U} \- \text{Set} \), the quadruple

\[
\left( Q(V), Q(A), Q(h), Q(t) \right)
\]

is a quiver in the sense of definition 1320.

No object in \( \mathcal{U} \- \text{Quiv} \) is formally equal to any object in \( [I, \mathcal{U} \- \text{Set}] \) in the sense of ZFC. They are, however, equivalent, as shown above, and this can be formalized by stating that the two categories are isomorphic, in the sense of definition 1125 (e), as objects of the category \( \mathcal{V} \- \text{Cat} \), where \( \mathcal{V} \) is a Grothendieck universe that strictly contains \( \mathcal{U} \). We have already defined this isomorphism explicitly.

This is an example of isomorphism of categories. In practice, if two categories are not so obviously identical, we are usually better served by equivalences of categories defined in definition 1169.

Definition 1160. The opposite natural transformation of \( \alpha : F \Rightarrow G \), where \( F \) and \( G \) are functors from \( \mathcal{C} \) to \( \mathcal{D} \), is the natural transformation \( \alpha^{\text{op}} : G^{\text{op}} \Rightarrow F^{\text{op}} \), in which we take the opposite of each component in \( \alpha \).

Dual natural transformation arise naturally in proposition 1161.

Proof of correctness. The naturality diagram (381) commutes for \( \alpha^{\text{op}} \) because all morphisms are simply reversed. \( \square \)

Proposition 1161. For the opposite of the functor category \( [\mathcal{C}, \mathcal{D}] \) we have

\[
[\mathcal{C}, \mathcal{D}]^{\text{op}} = [\mathcal{C}^{\text{op}}, \mathcal{D}^{\text{op}}].
\]

This is part of the duality principles listed in proposition 1124.

Proof. In definition 1145, we have defined the opposite functor \( F^{\text{op}} : \mathcal{C}^{\text{op}} \to \mathcal{D}^{\text{op}} \) of \( F : \mathcal{C} \to \mathcal{D} \) in a way that allows us to regard it as an object of \( [\mathcal{C}^{\text{op}}, \mathcal{D}^{\text{op}}] \).

In definition 1160, we have defined the opposite natural transformation \( \alpha^{\text{op}} : G^{\text{op}} \Rightarrow F^{\text{op}} \) of \( \alpha : F \Rightarrow G \) in a way that allows us to regard it as a morphism of \( [\mathcal{C}^{\text{op}}, \mathcal{D}^{\text{op}}] \).

Furthermore, \( \alpha^{\text{op}} : G^{\text{op}} \to F^{\text{op}} \) reverses the direction of its morphisms, and hence it is the dual to \( \alpha \) in the category \( [\mathcal{C}, \mathcal{D}]^{\text{op}} \). \( \square \)

Definition 1162. Given an index category \( I \) and an arbitrary category \( \mathcal{C} \), for any object \( A \)}
in C, we can define the **constant functor**

\[ \Delta^I_A : I \to C, \]

\[ \Delta^I_A(X) := X, \]

\[ \Delta^I_A(g : X \to Y) := \text{id}_A. \]

Given two objects A and B in C, a natural transformation \( \alpha : \Delta^I_A \Rightarrow \Delta^I_B \) is an **indexed family** that gives the same morphism for every object of the index category I.

Indeed, the diagram (381) in this case becomes

\[
\begin{array}{ccc}
A & \xrightarrow{\Delta^I_A(f) = \text{id}_A} & A \\
\downarrow{\alpha_A} & & \downarrow{\alpha_B} \\
B & \xrightarrow{\Delta^I_B(f) = \text{id}_B} & B
\end{array}
\]

(390)

This diagram implies that \( \alpha_A = \alpha_B \) for any two objects A and B in I. Therefore, all components of \( \alpha \) are equal to some morphism in \( C(A,B) \).

We can now define the I-shaped **diagonal functor** on C

\[ \Delta^I : C \to [I, C], \]

\[ \Delta^I(A) := \Delta^I_A, \]

\[ \Delta^I(f : A \to B) := \{f : A \to B\}_{k \in I}. \]

It is called a diagonal functor because, if I is a discrete category of two objects, then \( \Delta^I \) gives the diagonal of the product category \( C^2 \) by providing, for each object A of C, the ordered pair \( (A, A) \) (and similarly for morphisms).

**Proposition 1163.** Let \( F \) and \( G \) be parallel functors from the category C to D. The family (380) is an isomorphism in the corresponding functor category \([C, D]\) if and only if all of its components are isomorphisms and, for any morphism \( f : A \to B \) in C, the following diagram commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{F(f)} & F(B) \\
\downarrow{\alpha_A} & & \downarrow{\alpha_B^{-1}} \\
G(A) & \xrightarrow{G(f)} & G(B)
\end{array}
\]

(391)

We say that \( \alpha \) is a **natural isomorphism**.

**Proof.** If all components of \( \alpha \) are isomorphisms, the condition

\[ \alpha_B \circ F(f) = G(f) \circ \alpha_A \]
is equivalent to
\[ F(f) = \alpha_A^{-1} \circ G(f) \circ \alpha_A. \]

We must now show that, if \( \alpha \) is an isomorphism in \([C, D]\), all of its components are isomorphisms.

If \( \alpha : F \Rightarrow G \) is an isomorphism in \([C, D]\). Then there exists some natural transformation \( \beta : G \Rightarrow F \) such that
\[ \beta \circ \alpha = \text{id}_F \quad \text{and} \quad \alpha \circ \beta = \text{id}_G. \]

For every object \( A \) in \( C \), the morphism \( \alpha_A : F(A) \to G(A) \) composed with \( \beta_A : G(A) \to F(A) \) is
\[ \beta_A \circ \alpha_A = \text{id}_{F(A)}. \]
Therefore, \( \alpha_A \) is left-invertible. Analogously,
\[ \alpha_A \circ \beta_A = \text{id}_{F(A)} \]
and hence \( \alpha_A \) is right-invertible.

Therefore, for every object \( A \) in \( C \), the morphism \( \alpha_A \) is fully invertible, i.e. an isomorphism. \( \square \)

**Definition 1164.** We define the **product category** \( C \times D \) of \( C \) and \( D \) as follows:

- The **set of objects** is the Cartesian product
  \[ \text{obj}(C \times D) := \text{obj}(C) \times \text{obj}(D). \]  
  (392)

- The **set of morphisms** from the pair of objects \((A, X)\) to \((B, Y)\) is the product
  \[ (C \times D)((A, X), (B, Y)) := C(A, B) \times D(X, Y). \]  
  (393)

- The **composition of the morphisms**
  \[ (f, r) : (A, X) \to (B, Y) \]
  \[ (g, s) : (B, Y) \to (C, Z) \]
  is the pairwise composition
  \[ (g, s) \circ (f, r) := (g \circ f, s \circ r) \quad (A, X) \to (C, Z) \]  
  (394)
- The **identity morphism** of the pair \((A, X)\) is simply the pair of identity morphisms \((\text{id}_A, \text{id}_X)\).

**Definition 1165.** Let \( C \) be a locally \( U \)-small category. We can regard the morphism sets \( C(A, B) \) as a functor parameterized by objects of \( C \).
(a) For any pair of morphisms \( f : B \to A \) and \( g : X \to Y \) in \( \mathbf{C} \), define the operator
\[
T_{f,g} : \mathbf{C}(A,X) \to \mathbf{C}(B,Y)
\]
\[
T_{f,g}(s) \mapsto g \circ s \circ f.
\] (395)

The action of \( T_{f,g} \) can be expressed graphically as
\[
\begin{array}{c}
A \xrightarrow{f} X \\
B \xrightarrow{T_{f,g}(s)} Y
\end{array}
\]
(396)

We can now define the following **binary hom-functor**:
\[
\mathbf{C}(-,-) : \mathbf{C}^{\text{op}} \times \mathbf{C} \to \mathcal{U}_{-}\text{Set}
\]
\[
\mathbf{C}(A,X) := \{ s : A \to X \}
\]
\[
\mathbf{C}(f,g) := T_{f,g}
\] (397)

(b) Fixing the first argument \( A \) in (397), we instead obtain a covariant unary hom-functor:
\[
\mathbf{C}(A,-) : \mathbf{C} \to \mathcal{U}_{-}\text{Set}
\] (398)

Analogously, fixing the second argument \( X \), we obtain a **contravariant** unary hom-functor:
\[
\mathbf{C}(-,X) : \mathbf{C}^{\text{op}} \to \mathcal{U}_{-}\text{Set}
\] (399)

**Proof of correctness.** It is sufficient to verify that (397) defines a functor. CF2 can be seen to hold by inspecting the diagram:
\[
\begin{array}{c}
A \xrightarrow{f} X \\
B \xrightarrow{T_{f_1,s_1}(s)} Y \\
C \xrightarrow{\left[T_{f_2,\delta_2} \circ T_{f_1,s_1}\right](s)} Z
\end{array}
\]
(400)

The other functor condition CF1 is straightforward to prove. \( \square \)

**Proposition 1166.** **Function currying** is a natural isomorphism between the functors
\[
\text{Set}(A \times B, C) \quad \quad \quad \text{Set}(A, \text{Set}(B, C))
\]

More concretely, consider the following functors, which are loosely based on definition 1165:
\[
V : \text{Set}^3 \to \text{Set}
\]
\[
V(A, B, C) := \text{Set}(A \times B, C)
\]
\[
\left(V(f : X \to A, g : Y \to B, h : C \to Z)\right)(s : A \times B \to C) := (x,y) \mapsto h\left(s\left(f(x), g(y)\right)\right)
\]

592
and

\[ W : \text{Set}^3 \rightarrow \text{Set} \]

\[ W(A, B, C) := \text{Set}(A, \text{Set}(B, C)) \]

\[ W(f : X \rightarrow A, g : Y \rightarrow B, h : C \rightarrow Z)(t : A \rightarrow \text{Set}(B, C)) := x \mapsto y \mapsto h\left(t(f(x))(g(y))\right) \]

Then the family of functions

\[ \alpha : V \Rightarrow W \]

\[ \alpha_{A,B,C}(s : A \times B \rightarrow C) := a \mapsto b \mapsto s(a, b) \]

is a natural isomorphism.

**Proof.** The function \( \varphi \) is clearly invertible. Fix a triple of functions \( f : X \rightarrow A, g : Y \rightarrow B \) and \( h : C \rightarrow Z \). For every \( s : A \times B \rightarrow C \) we have

\[ [W(f, g, h)](\alpha_{A,B,C}(s)) = x \mapsto y \mapsto \left(h\left([a \mapsto b \mapsto s(a, b)](f(x))(g(y))\right)\right) = \\
= x \mapsto y \mapsto h\left(s(f(x), g(y))\right) = \\
= \alpha_{X,Y,Z}(V(f, g, h)), \]

which proves that the following diagram commutes:

\[
\begin{array}{ccc}
V(A, B, C) & \xrightarrow{V(f,g,h)} & V(X, Y, Z) \\
\downarrow{\alpha_{A,B,C}} & & \downarrow{\alpha_{X,Y,Z}} \\
W(A, B, C) & \xrightarrow{W(f,g,h)} & W(X, Y, Z)
\end{array}
\]

(401)
14.3. Category equivalences

Remark 1167. We have the following notions for expressing that two categories \( \mathcal{C} \) and \( \mathcal{D} \) are similar:

(a) Obviously, if \( \mathcal{C} \) and \( \mathcal{D} \) are equal, they are similar.

(b) A slightly less obvious notion is isomorphism of categories. This is an isomorphism, in the sense of definition 1125 (e), in the category \( \mathcal{U} \text{-}\mathbf{Cat} \) of small categories for a suitable Grothendieck universe \( \mathcal{U} \). That is, \( \mathcal{C} \) and \( \mathcal{D} \) are isomorphic if there exists an invertible functor between them.

We rarely distinguish between objects and arrows of isomorphic categories, even if we do not have strict equality in the sense of the axiom of extensionality in ZFC.

Examples of isomorphic categories include theorem 1186 (Ordered sets as categories) and example 1159.

(c) A weaker but very useful notion is category equivalence defined in definition 1169.

Example 1168. There is an isomorphism between the category \( \mathcal{U} \text{-}\mathbf{Set} \) of small sets and \( \mathcal{U} \text{-}\mathbf{DiscrCat} \) of all small discrete categories.

Consider the pair of functors

\[
U : \mathcal{U} \text{-}\mathbf{DiscrCat} \to \mathcal{U} \text{-}\mathbf{Set},
\]

which for any small category \( \mathcal{C} \) gives us its set of objects \( \text{obj}(\mathcal{C}) \) and

\[
D : \mathcal{U} \text{-}\mathbf{Set} \to \mathcal{U} \text{-}\mathbf{DiscrCat},
\]

which for any small set \( A \) gives us the discrete category whose set of objects is \( A \).

These were discussed in example 1141, although with \( \mathbf{Cat} \) rather than \( \mathcal{U} \text{-}\mathbf{DiscrCat} \).

It is obvious that for any small category \( \mathcal{C} \), the functor \( D \circ U \) is bijective on objects. We need to verify that it is bijective on morphisms, however, in order to prove that \( D \circ U \) is the identity functor on \( \mathcal{U} \text{-}\mathbf{DiscrCat} \).

But any functor \( F : \mathcal{C} \to \mathcal{D} \) is completely determined by how \( F \) acts on the objects of \( \mathcal{C} \). Indeed, the only morphisms in \( \mathcal{C} \) are the identity morphisms and, by CF1, every object \( X \) in \( \mathcal{C} \) determines how the identity \( \text{id}_X \in C(X) \) is mapped by \( F \).

Therefore, \( D \) is a left inverse of \( U \). It is also a right inverse — for any function \( f : A \to B \) between small sets,

\[
[U \circ D](f) = D(f)|_A = f.
\]

Therefore, the forgetful functor \( U \) is invertible, and its inverse is \( D \).

Definition 1169. An equivalence between the categories \( \mathcal{C} \) and \( \mathcal{D} \) is a quadruple

\[
\begin{align*}
F &: \mathcal{C} \to \mathcal{D}, \\
G &: \mathcal{D} \to \mathcal{C}, \\
\eta &: \text{id}_C \Rightarrow G \circ F, \\
\varepsilon &: F \circ G \Rightarrow \text{id}_D,
\end{align*}
\]

(402)

[Lei16] def. 1.3.15
where $\eta$ and $\varepsilon$ are natural isomorphisms.

We call $\eta$ the **unit** of the equivalence and $\varepsilon$ the **counit**.

If $(F, G, \eta, \varepsilon)$ is an equivalence, we say that $\mathcal{C}$ and $\mathcal{D}$ are equivalent categories. This is justified because equivalence of categories is an equivalence relation — see proposition 1170.

Note that an equivalence is not necessarily an adjunction, they simply have a common setup. This is discussed in proposition 1194.

**Proposition 1170.** For every Grothendieck universe $\mathcal{U}$, category equivalence is an equivalence relation on the set $\text{obj}(\mathcal{U}\text{-Cat})$

*Proof.*

**Proof of reflexivity.** Clearly $(\text{id}_\mathcal{C}, \text{id}_\mathcal{C}, \text{id}_{\text{id}_\mathcal{C}}, \text{id}_{\text{id}_\mathcal{C}})$ is a self-equivalence for the category $\mathcal{C}$.

**Proof of symmetry.** If $(F, G, \eta, \varepsilon)$ is an equivalence between the categories $\mathcal{C}$ and $\mathcal{D}$, then $(G, F, \varepsilon^{-1}, \eta^{-1})$ is an equivalence between $\mathcal{D}$ and $\mathcal{C}$.

**Proof of transitivity.** By proposition 1128 (i), the composition of invertible natural isomorphisms is again a natural isomorphism, hence equivalence in $\text{obj}(\mathcal{U}\text{-Cat})$ is a transitive relation.

**Proposition 1171.** Let $\mathcal{C}$ and $\mathcal{D}$ be discrete categories. Then $\mathcal{C}$ and $\mathcal{D}$ are equivalent if and only if the underlying sets $\text{obj}(\mathcal{C})$ and $\text{obj}(\mathcal{D})$ are equinumerous.

*Proof.*

**Proof of sufficiency.** Suppose that $(F, G, \eta, \varepsilon)$ be a category equivalence.

The unit natural transformation $\eta : \text{id}_\mathcal{C} \Rightarrow G \circ F$ consists of a morphism

$$\eta_A : A \to (G \circ F)(A)$$

for every object $A$ of $\mathcal{C}$. Since the only morphisms in $\mathcal{C}$ are the identities, it follows that $\eta_A = \text{id}_A$ and hence $(G \circ F)(A) = A$. In particular, this implies that $\eta$ is the identity natural transformation on $\text{id}_\mathcal{C}$ and that the restriction $G|_{\text{obj}(\mathcal{D})}$ is a left inverse of $F|_{\text{obj}(\mathcal{C})}$.

Similarly, for the counit $\varepsilon : F \circ G \Rightarrow \text{id}_\mathcal{D}$, for every object $X$ in $\mathcal{D}$ we have $\varepsilon_X = \text{id}_X$ and hence $(F \circ G)(X) = X$. Thus, $\eta$ is the identity natural transformation on $\text{id}_\mathcal{D}$ and $G|_{\text{obj}(\mathcal{D})}$ is a right inverse of $F|_{\text{obj}(\mathcal{C})}$.

Therefore, the sets $\text{obj}(\mathcal{C})$ and $\text{obj}(\mathcal{D})$ are equinumerous.

**Proof of necessity.** Suppose that $F : \text{obj}(\mathcal{C}) \to \text{obj}(\mathcal{D})$ is a bijective function. Then it is an isomorphism in the category $\mathcal{U}\text{-Cat}$ for an appropriate universe $\mathcal{U}$, hence it induces an equivalence between $\mathcal{C}$ and $\mathcal{D}$.

**Proposition 1172.** The opposite of equivalent categories are equivalent.

More precisely, if $(F, G, \eta, \varepsilon)$ is an equivalence between the categories $\mathcal{C}$ and $\mathcal{D}$, then

$$G^\text{op} : \mathcal{D}^\text{op} \to \mathcal{C}^\text{op},$$

$$F^\text{op} : \mathcal{C}^\text{op} \to \mathcal{D}^\text{op},$$

$$\varepsilon^\text{op} : \text{id}_\mathcal{D} \Rightarrow (F \circ G)^\text{op},$$

$$\eta^\text{op} : (G \circ F)^\text{op} \Rightarrow \text{id}_{\mathcal{C}^\text{op}},$$

$$G^\text{op} \circ F^\text{op}$$

595
is an equivalence between $\mathbf{C}^{\text{op}}$ and $\mathbf{D}^{\text{op}}$.

This is part of the duality principles listed in proposition 1124.

Proof. Trivial. $\square$

**Proposition 1173.** In any category equivalence $(F, G, \eta, \varepsilon)$, the functor $F$ is fully faithful and essentially surjective on objects.

The converse of this statement is theorem 1175 (Fully faithful and essentially surjective functor induces equivalence).

Proof.

**Proof of essential surjectivity.** For any object $X$ in $\mathbf{D}$, $A := G(X)$ is an object in $\mathbf{C}$.

By definition of category equivalence, the morphism

$$\varepsilon_X : \underbrace{F \circ G}_F(A) \to X$$

is an isomorphism.

Therefore, for every object $X$ in $\mathbf{D}$, there exists some object $A$ in $\mathbf{C}$ such that $F(A) \cong X$.

Thus, $F$ is essentially surjective.

**Proof of faithfulness.** Fix some objects $A$ and $B$ in $\mathbf{C}$. Let $f_1 : A \to B$ and $f_2 : A \to B$ be morphisms such that $F(f_1) = F(f_2)$.

From the naturality of $\eta$ it follows that the following diagram commutes:

$$
\begin{array}{ccc}
A & \xrightarrow{f_1} & B \\
\eta_A \downarrow & & \downarrow \eta_B \\
[G \circ F](A) & \xrightarrow{[G \circ F](f_1)} & [G \circ F](B) \\
\eta_A \downarrow & & \downarrow \eta_B \\
A & \xrightarrow{f_2} & B
\end{array}
$$

Therefore, $\eta_B \circ f_1 = \eta_B \circ f_2$ and, since $\eta_B$ is left-cancellative, $f_1 = f_2$.

**Proof of fullness.** Fix some objects $A$ and $B$ in $\mathbf{C}$. Let $g : F(A) \to F(B)$ be an arbitrary morphism.

We can define a morphism $f : A \to B$ via the composition

$$
\begin{array}{ccc}
A & \xrightarrow{f} & B \\
\eta_A \downarrow & \quad & \downarrow \eta_B \\
[G \circ F](A) & \xrightarrow{G(g)} & [G \circ F](B)
\end{array}
$$

(405)
By naturality of $\eta$, the following diagram commutes:

\[
\begin{array}{ccc}
G \circ f(A) & \xrightarrow{G(g)} & G \circ F(B) \\
\eta_A & \downarrow & \eta_B \\
A & \xrightarrow{f} & B \\
\eta_A^{-1} & \uparrow & \eta_B^{-1} \\
G \circ f(A) & \xleftarrow{G \circ F(f)} & G \circ F(B)
\end{array}
\]

(405)

Therefore,

\[G(g) = [G \circ F](f).\]

We have already shown that $F$ is faithful and, by proposition 1170, $G$ is also faithful. Since $g$ and $F(f)$ are parallel, it follows that they are equal.

Therefore, $F$ is full. \qed

**Remark** 1174. In the fullness proof of proposition 1173, we can use another argument if $(F, G, \eta, \varepsilon)$ is an adjoint equivalence.

If the triangle diagram (418) commutes, the dashed lines in the following diagram also commute:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{F(g)} & F(B) \\
\eta_A & \downarrow & \eta_B \\
F(\eta_A(F)) & \xrightarrow{[F \circ G](g)} & F \circ G \circ F(A) \\
\varepsilon_{F(A)} & \downarrow & \varepsilon_{F(B)} \\
F(A) & \xleftarrow{g} & F(B)
\end{array}
\]

(406)

**Theorem 1175** (Fully faithful and essentially surjective functor induces equivalence). *Every fully faithful and essentially surjective on objects functor induces a category equivalence.*

More precisely, given a functor $F : \mathcal{C} \to \mathcal{D}$ that is fully faithful and essentially surjective on objects, there exists a functor $G : \mathcal{D} \to \mathcal{C}$ and natural transformations

\[\eta : \text{id}_\mathcal{C} \Rightarrow G \circ F,\]

\[\varepsilon : F \circ G \Rightarrow \text{id}_\mathcal{D},\]

such that the quadruple $(F, G, \eta, \varepsilon)$ is a category equivalence.

In ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (h).

We prove the converse of this statement separately in proposition 1173.

**Proof.**

**Proof that the axiom of choice implies functors induce equivalences.** Suppose that the axiom of choice holds and let $F$ be a fully faithful functor that is surjective on objects.

From essential surjectivity of $F$, it follows that for every object $X$ in $\mathcal{D}$, the preimage of $X$ under $F$ is nonempty. The preimage of $X$ is the set $\mathcal{A}_X$ of objects in $\mathcal{C}$ such that $A \in \mathcal{A}_X$ if
and only if \( F(A) \cong X \). We use the axiom of choice on the family \( \{ A_X \}_{X \in \mathcal{D}} \) to select a single preimage for every \( X \), which we denote by \( G(X) \).

Again using the axiom of choice, we pick an isomorphism \( \varepsilon_X : F(G(X)) \rightarrow X \) for every \( X \).

We have defined a function \( G \) from \( \text{obj}(\mathcal{D}) \) to \( \text{obj}(\mathcal{C}) \). In order to \( G \) to become a functor, we must extend it to morphisms. Let \( X \) and \( Y \) be objects in \( \mathcal{D} \) and \( g : X \rightarrow Y \) be any morphism.

Consider the morphism

\[
\varepsilon_Z^{-1} \circ f \circ \varepsilon_X : [F \circ G](X) \rightarrow [F \circ G](Y).
\]

Since \( F \) is fully faithful, there exists a unique morphism \( g \) in \( \mathcal{D}(G(X), G(Y)) \) such that the following diagram commutes:

\[
\begin{array}{ccc}
[F \circ G](X) & \xrightarrow{F(f)} & [F \circ G](Z) \\
\varepsilon_X & & \varepsilon'_Z \\
X & \xrightarrow{g} & Z
\end{array}
\]

(407)

We define \( G(g) := g \).

In order to prove that \( G \) is a functor, we need to show that \( CF1 \) and \( CF2 \) hold.

For \( CF1 \), note that the following diagram commutes for any object \( X \) in \( \mathcal{D} \):

\[
\begin{array}{ccc}
[F \circ G](X) & \xrightarrow{[F \circ G](\text{id}_X)} & [F \circ G](X) \\
\varepsilon_X & & \varepsilon'_X \\
X & \xrightarrow{\text{id}_X} & X
\end{array}
\]

(408)

Note that (407) also commutes if we replace \( [F \circ G](\text{id}_X) \) with \( F(\text{id}_{G(X)}) \). Since \( F \) is fully faithful, this morphism is unique and it follows that

\[
[F \circ G](\text{id}_X) = F(\text{id}_{G(X)}).
\]

For \( CF2 \), analogously, given morphisms \( g : X \rightarrow Y \) and \( f : Y \rightarrow Z \), the following diagram commutes:

\[
\begin{array}{ccc}
[F \circ G](X) & \xrightarrow{[F \circ G](f)} & [F \circ G](Z) \\
\varepsilon_X & & \varepsilon'_Z \\
X & \xrightarrow{g} & Z \\
\varepsilon_Z & & \varepsilon'_Z \\
Z & \xrightarrow{f} & U
\end{array}
\]

(409)

We have implicitly used remark 1149 above.

By the same uniqueness argument used for \( CF1 \), we conclude that

\[
G(f \circ g) = G(f) \circ G(g).
\]

598
We have shown that $G$ is a functor. Furthermore, $\varepsilon$ is a natural transformation since, for any morphism $g : X \to Y$ in $D$, the following diagram commutes:

\[
\begin{array}{c}
X \\
\downarrow \varepsilon_x \\
F \circ G(X) \\
\downarrow (F \circ G)(g) \\
\varepsilon^\sim \\
F \circ G(X) \\
\downarrow \varepsilon^\sim_Z \\
Z \\
\end{array}
\]

To show that $F$ induces an equivalence, it now only remains to define a unit natural transformation $\eta : \text{id}_C \Rightarrow G \circ F$. For every object $A$ in $C$ we have an isomorphism

\[
\varepsilon^{-1}_{F(A)} : F(A) \to [F \circ G \circ F](A).
\]

Using $G(\varepsilon^{-1}_{F(A)})$ will get us nowhere. Fortunately, $F$ is fully faithful, so there is a bijective function

\[
\varphi : D(F(A), [F \circ G \circ F](A)) \to C(A, [F \circ G](A)).
\]

Hence, we can define

\[
\eta : \text{id}_C \Rightarrow F \circ G
\]

\[
\eta_A := \varphi(\varepsilon^{-1}_{F(A)})
\]

so that $F(\eta_A) = \varepsilon^{-1}_{F(A)}$. By proposition 1152 (h), since $\varepsilon_{F(A)}$ is an isomorphism, $\eta_A$ is also an isomorphism.

By naturality of $F(\eta_A)$, the following diagram commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{F(f)} & F(B) \\
\downarrow F(\eta_A) & & \downarrow F(\eta_B) \\
[F \circ G \circ F](A) & \xrightarrow{[F \circ G \circ F](f)} & [F \circ G \circ F](B) \\
\end{array}
\]

Hence, by proposition 1151, the following diagram also commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{[G \circ F](f)} & B \\
\downarrow \eta_A & & \downarrow \eta_B \\
[G \circ F](A) & \xrightarrow{[G \circ F](f)} & [G \circ F](B) \\
\end{array}
\]

Therefore, the quadruple $(F, G, \eta, \varepsilon)$ is an equivalence of categories.

**Proof that functors induce equivalences implies the axiom of choice.** Let $\mathcal{A}$ be a family of nonempty sets. Let $D$ be the discrete category induced by $\mathcal{A}$.

Define the category $C$ as follows:

- The set of objects $\text{obj}(C)$ is the disjoint union $\bigsqcup_{A \in \mathcal{A}} A$.
- The set of morphisms $C((A, x), (B, y))$ has a single morphism if $A = B$ and no morphisms otherwise. This single morphism can be encoded as the triple $(A, x, y)$.
The composition of the morphisms \((A, x, y)\) and \((A, y, z)\) is the morphism \((A, x, z)\).

The identity morphism on the object \((A, x) \in C\) is \((A, x, x)\).

Define the functor

\[
F : C \to D
\]

\[
F(A, x) := A
\]

\[
F(A, x, y) := \text{id}_A
\]

that maps each point \(x \in A \in \mathcal{A}\) into the set \(A\) it belongs to. We have taken the disjoint union of \(\mathcal{A}\) since otherwise there may not be a canonical choice of set \(A\) for \(F\) to send \(x\) to. Thus, the functor is surjective on objects (not essentially surjective but actually surjective).

Note that \(D(F(A, x), F(B, y))\) has a single morphism if \(A = B\) and is empty otherwise. From this it follows that \(F\) is fully faithful.

Therefore, \(F\) induces a category equivalence \((F, G, \eta, \varepsilon)\). The functor \(G\) chooses an object \((A, x)\) of \(C\) for each object \(A\) of \(D\). This induces a choice function on \(\mathcal{A}\).

We have shown that the axiom of choice holds. \(\Box\)

**Definition 1176.** A groupoid is a category whose only morphisms are isomorphisms.

**Definition 1177.** Let \((M, \cdot, e)\) be a monoid. The delooping \(B_M\) of \(M\) is the following category:

- The set of objects \(\text{obj}(B_M)\) is the singleton set \{\(\bullet\)\}, where \(\bullet\) is any set not in the set of all morphisms \(M\).
- The only set of morphisms \(B_M(\bullet)\) is the underlying set \(M\) of the monoid.
- The composition of the morphisms \(x\) and \(y\) is the multiplication:

\[
y \circ x := y \cdot x.
\]

Note how we write composition in the same order as multiplication. This may seem to contradict the general convention, however it is consistent with groups being regarded as sets of invertible transformations.

- The identity morphism is \(e\).

**Proposition 1178.** The delooping of a group \(G\) is a groupoid.

There is a restricted form of a converse — see proposition 1180.

**Proof.** Trivial. \(\Box\)

**Definition 1179.** A category is connected if there exists a morphism between any two objects. That is, for a connected category \(C\), given two objects \(X\) and \(Y\), either \(C(X, Y)\) or \(C(Y, X)\) is nonempty.

Connected categories are used in proposition 1180 and theorem 1186 (c).
Proposition 1180. For every connected groupoid $G$ there exists a group $G$ such that $G$ is equivalent to the delooping $BG$. Furthermore, if $G$ has only one object, then this equivalence is an isomorphism.

See proposition 1178 for a much simpler converse.

Proof. Define the group $G$ as follows:

- Let the underlying set of $G$ be the set of all morphisms of $G$.
- Define $x \cdot y = z$ to hold whenever the following diagram commutes:

\[
\begin{array}{ccc}
\bullet & \rightarrow & \bullet \\
\downarrow & & \downarrow \\
\bullet & \rightarrow & \bullet \\
\end{array}
\]

- Pick the identity of $G$ out of the identity morphisms in $G$.
- Let the inverse of $x$ be the inverse morphism $x^{-1}$.

Finally, define the functor $F : G \to BG$, which sends every object to $\bullet$ and every morphism to itself.

By definition, for every morphism $x : \bullet \to \bullet$ in the delooping, there exists a unique $f : X \to Y$ in $G$ such that $F(f) = x$. Hence, $f$ is surjective on morphisms, and by proposition 1152 (b), full and surjective on objects.

Furthermore, for any objects $X$ and $Y$ in $G$, we have

\[G(X, Y) \subseteq BG(F(X), F(Y)).\]

Hence, $G$ is faithful. Therefore, by theorem 1175 (Fully faithful and essentially surjective functor induces equivalence), the groupoid $G$ is equivalent to the delooping $BG$.

Furthermore, if $G$ has only one object, it is also injective on objects and, by proposition 1152 (a), injective on morphisms. In this case, we have an isomorphism due to proposition 1152 (i). \qed

Definition 1181. The category $S$ is called skeletal if the only isomorphisms in $S$ are the identity morphisms.

If $S$ is a subcategory of $C$ and if they are equivalent, we say that $S$ is a skeleton of $C$.

Example 1182. Fix a Grothendieck universe $\mathcal{U}$, the cardinal numbers in $\mathcal{U}$ are a skeletal subcategory of $\mathcal{U}$-$\text{Set}$.

Theorem 1183 (Category skeleton existence). Every category has a skeleton.

In ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (i).

Proof.
Proof that the axiom of choice implies skeleton existence. Suppose that the axiom of choice holds and let $F$ be a fully faithful functor that is surjective on objects.

Fix a category $\mathcal{C}$. We will build a subcategory $\mathcal{S}$ of $\mathcal{C}$ whose inclusion functor $I : \mathcal{S} \rightarrow \mathcal{C}$ is essentially surjective and fully faithful. By theorem 1175 (Fully faithful and essentially surjective functor induces equivalence), this is sufficient for $\mathcal{S}$ and $\mathcal{C}$ to be equivalent.

In order for $I$ to be a full functor, $\mathcal{S}$ must be a full subcategory. Therefore, when building $\mathcal{S}$, we can only remove objects and must preserve the morphism sets for the remaining objects.

Denote by obj($\mathcal{C}$) $\cong$ the quotient of $\mathcal{C}$ by the isomorphism relation. Using the axiom of choice, we can obtain a choice function $c : \text{obj}($\mathcal{C}$) $\cong$ $\rightarrow$ $\text{obj}($\mathcal{C}$)$.

Define $\mathcal{S}$ as the subcategory induced by the image $c[\text{obj}($\mathcal{C}$) $\cong$].

Now consider the inclusion functor $I : \mathcal{S} \rightarrow \mathcal{C}$. For every pair $X$ and $Y$ of objects in $\mathcal{S}$, clearly $\mathcal{S}(X, Y) = \mathcal{C}(I(X), I(Y))$.

Hence, $I$ is fully faithful.

Now let $X$ and $Y$ be objects of $\mathcal{C}$. Since the objects of $\mathcal{S}$ were chosen from isomorphism classes of $\mathcal{C}$, there exist objects $X'$ and $Y'$ in $\mathcal{S}$ that are isomorphic to $X$ and $Y$, correspondingly. Hence, $I$ is essentially surjective.

Therefore, $F$ satisfies theorem 1175 (Fully faithful and essentially surjective functor induces equivalence), from which it follows that $\mathcal{C}$ and $\mathcal{S}$ are equivalent.

Proof that skeleton existence implies the axiom of choice. Suppose that every category has a skeleton.

Let $\mathcal{A}$ be a family of nonempty sets. Construct a category $\mathcal{C}$ from the disjoint union $\bigsqcup_{A \in \mathcal{A}} A$, where a morphism exists only between members of the same set. This construction is performed in detail in the proof of theorem 1175 (Fully faithful and essentially surjective functor induces equivalence).

Then $\mathcal{C}$ has a skeleton $\mathcal{S}$. All morphisms in $\mathcal{C}$ are isomorphisms, hence the set $\text{obj}(\mathcal{S})$ contains exactly one representative for each set in the family $\mathcal{A}$.

More precisely, define the set

$S := \{x \mid (A, x) \in \text{obj}(\mathcal{S})\}$.

Then $S$ satisfies theorem 990 (a).

Since the family $\mathcal{A}$ is arbitrary, we conclude that the axiom of choice holds. □

Remark 1184. Rather than defining representatives of equivalence classes, as in theorem 1183 (Category skeleton existence), we can define morphisms between the equivalence classes themselves, as in proposition 1226.

This does not require the axiom of choice, but is rarely applicable, unfortunately. One case where it is applicable is in thin categories — see theorem 1186 (Ordered sets as categories).

[nLa20] Definition 1185. A category is thin if any two parallel morphisms are equal.

This is equivalent to saying that the function for every two objects $A$ and $B$ in $\mathcal{P}$, whenever the set $\mathcal{P}(A, B)$ is at most a singleton.
As shown in example 1225 and discussed in theorem 1186 (Ordered sets as categories), a thin category may not be skeletal.

Thin categories are often conflated with preordered sets due to theorem 1186 (a).

**Theorem 1186** (Ordered sets as categories). Regarding ordered sets, we have the following category isomorphisms:

(a) The categories **PreOrd** and **Thin**, where
   - **PreOrd** is the category of small preordered sets and (nonstrict) monotone maps.
   - **Thin** is the subcategory of **Cat** induced by thin categories, i.e. the category of small thin categories.

(b) The categories **Pos** and **SkelThin**, where
   - **Pos** is the subcategory of **PreOrd** induced by (nonstrict) partially ordered sets.
   - **SkelThin** is the subcategory of **Thin** induced by skeletal categories, i.e. the category of small thin skeletal categories.

(c) The categories **Tos** and **ConnSkelThin**, where
   - **Tos** is the subcategory of **Pos** induced by (nonstrict) totally ordered sets.
   - **ConnSkelThin** is the subcategory of **SkelThin** induced by connected categories, i.e. the category of small thin skeletal connected categories.

**Proof.**

**Proof of 1186 (a).**

**Proof that preorders induce thin categories.** Let \((P, \leq)\) be a small preordered set. Let \(P\) be the free category obtained by regarding \((P, \leq)\) as a quiver. Explicitly, the category \(P\) is built as follows:

- The set of objects \(\text{obj}(P)\) is simply \(P\).
- The set of morphisms \(P(x, y)\) consists of the tuple \((x, y)\) if \(x \leq y\) and is empty otherwise.
- The composition of the morphisms \((x, y)\) and \((y, z)\) is simply \((x, y)\). This is well-defined because of the transitivity of \(\leq\).
- The identity morphism on the object \(x \in C\) is \((x, x)\). This is well-defined because of the reflexivity of \(\leq\).

The category is clearly thin because \(\leq\) is a binary relation and ordered tuples with the same elements are equal.

Now let \(f : P \rightarrow Q\) be a nonstrict monotone map. It induces the functor

\[
F : P \rightarrow Q
\]

\[
F(x) := f(x)
\]

\[
F(x, y) := (f(x), f(y)).
\]

CF1 is immediate and CF2 follows from (492), hence \(F\) is indeed a functor.
**Proof that thin categories induce preorders.** Let \( \mathcal{P} \) be a small category. Define the binary relation
\[
X \leq Y \text{ if and only if } \mathcal{P}(X, Y) \neq \emptyset.
\]
This is a binary relation over the set \( \mathcal{P} := \text{obj}(\mathcal{P}) \). It is reflexive because of the existence of identity morphisms in \( \mathcal{P} \) and transitive because of the requirement that the composition of compatible morphisms exists.

Therefore, \((\mathcal{P}, \leq)\) is a preordered set.

Given a functor \( F : \mathcal{P} \to \mathcal{Q} \), the restriction \( F|_{\text{obj}(\mathcal{P})} \) is a monotone map from \((\mathcal{P}, \leq_\mathcal{P})\) to \((\mathcal{Q}, \leq_\mathcal{Q})\).

Indeed, if \( X \leq Y \) for \( X, Y \in \mathcal{P} \), then \( \mathcal{P}(X, Y) \neq \emptyset \). Hence, \( \mathcal{P}(F(X), F(Y)) \neq \emptyset \) and \( F(X) \leq F(Y) \).

**Proof of isomorphism.** We have implicitly defined a functor from \textbf{PreOrd} to \textbf{Thin} and vice versa. Isomorphism of \textbf{PreOrd} and \textbf{Thin} requires that these functors are mutually inverse.

We need to prove that the induced preordered set for the induced thin category of a preordered set is the same as the original. In the other direction, we need to prove that the induced thin category for the induced preordered set of a thin category is the same as the original.

Both of these proofs are trivial but nevertheless the fact that we need to perform such a check is important.

**Proof of 1186 (b).** We have already shown in theorem 1186 (a) the isomorphism between \textbf{PreOrd} and \textbf{Thin}.

For a thin category, there is an isomorphism between \( x \) and \( y \) if and only if there is both a morphism from \( x \) to \( y \) and one from \( y \) to \( x \). This isomorphism may not be unique as a consequence of example 1225. Uniqueness requires \( x = y \) to hold in this case, which is in turn equivalent to partial order antisymmetry.

**Proof of 1186 (c).** We have already shown in theorem 1186 (b) the isomorphism between \textbf{Pos} and \textbf{SkelThin}.

Connectedness of a category \( \mathcal{P} \) is then equivalent to totality of \((\mathcal{P}, \leq)\). \( \square \)

**Proposition 1187.** Let \((\mathcal{P}, \leq)\) be a partially ordered set and let \( \mathcal{P} \) be its corresponding category induced by theorem 1186 (Ordered sets as categories).

(a) The opposite category \( \mathcal{P}^{op} \) corresponds to the opposite partially ordered set \((\mathcal{P}, \geq)\).

(b) If \( \mathcal{P} \) has an initial object, since it is a skeletal category, this initial object is unique.

An object \( I \) is an initial object if and only if it is the bottom of \( \mathcal{P} \).

Dually, \( T \) is a terminal object if and only if it is the top of \( \mathcal{P} \).

**Proof.** Trivial. \( \square \)
14.4. Category adjunctions

Remark 1188. Suppose we have the functors $F : C \to D$ and $G : D \to C$.

- If $F$ is a left inverse to $G$, given only $G(X)$, $F$ can restore $X$.
- If $F$ is instead a left adjoint to $G$, given $G(X)$ and some object $A$ in $C$, $F$ can give us an object $F(A)$ such that the morphisms from $F(A)$ to $X$ are uniquely defined by those from $A$ to $G(X)$. Thus, $F$ cannot restore $X$, but it can give us objects in $D$ that act on $X$ as their images under $G$ would act on $G(X)$ in $C$. This is described in [Mar19] as $F$ being a “conceptual inverse” of $G$.
- If $F$ is a right adjoint to $G$, the morphisms from $X$ to $F(A)$ are determined by those from $G(X)$ to $A$.

[UNDEFINED] contains two equivalent definition of an adjunction, and remark 1198 describes how they can be characterized via universal mapping properties.

Definition 1189. An adjunction between the categories $C$ and $D$ can be defined in several equivalent ways. Let $F : C \to D$ and $G : D \to C$ be arbitrary functors.

In both cases below, if there exists an adjunction between $F$ and $G$, we say that $F$ is left adjoint to $G$ and, correspondingly, that $G$ is right adjoint to $F$. A conventional notation for adjoint functors is $F \dashv G$.

(a) A hom-adjunction is a triple $(F, G, \varphi)$, where $\varphi$ is natural isomorphism

$$\varphi : D(F(-), -) \Rightarrow C(-, G(-)).$$

The functors

$$D(F(-), -) : C^{op} \times D \to \text{Set},$$
$$C(-, G(-)) : C^{op} \times D \to \text{Set}$$

are straightforward modifications of the binary hom-functor on $C$.

Naturality of $\varphi$ means that, for every two morphisms $f : B \to A$ in $C$ and $g : X \to Y$ in $D$, the following diagram commutes:

$$
\begin{array}{ccc}
D(F(A), X) & \xrightarrow{D(F(f)_X)} & D(F(B), Y) \\
\varphi_{A,X} \downarrow & & \downarrow \varphi_{B,Y} \\
C(A, G(X)) & \xrightarrow{C(f, G(g))} & C(B, G(Y))
\end{array}
$$

(415)

The above diagram reduces to the simpler to verify condition that, for every triple of morphisms $f : B \to A$, $s : F(A) \to X$ and $g : X \to Y$, we must have the equality

$$G(g) \circ \varphi_{A,X}(s) \circ f = \varphi_{B,Y}(g \circ s \circ F(f)).$$

(416)
(b) A unit-counit adjunction is a quadruple \((F, G, \eta, \varepsilon)\), where

\[
\eta : \text{id}_C \Rightarrow G \circ F, \\
\varepsilon : F \circ G \Rightarrow \text{id}_D
\]

are natural transformations satisfying the condition that, for any pair of objects \(A\) in \(C\) and \(Y\) in \(D\), the following triangle diagrams commute:

\[
\begin{align*}
F(A) & \xrightarrow{F(\eta_A)} F(A) \\
\eta_G(A) & \xrightarrow{\varepsilon_F(A)} [G \circ F](A)
\end{align*}
\]

\[
\begin{align*}
G(X) & \xrightarrow{G(\varepsilon_X)} G(X) \\
\eta_F(X) & \xrightarrow{\eta_G(X)} [G \circ F \circ G](X)
\end{align*}
\] (418) (419)

Note that an adjunction is not an equivalence, they simply have a common setup.

Similarly to equivalence, we call the natural transformation \(\eta\) the unit of the adjunction and \(\varepsilon\) the counit.

Proof of correctness.

Proof that 1189 (a) implies 1189 (b). Let \((F, G, \varphi)\) be a hom-adjunction.

For every morphism \(f : B \rightarrow A\) in \(C\), from the naturality of \(\varphi\) we have

\[
\begin{align*}
\text{D}(F(A), F(A)) & \xrightarrow{D(F(id_A), id_A)} \text{D}(F(B), F(A)) \\
\varphi_{A,F(A)} & \xrightarrow{\varphi_{A,F(A)}} \text{D}(F(B), F(B))
\end{align*}
\]

\[
\begin{align*}
\text{C}(A, [G \circ F](A)) & \xrightarrow{C_f([G \circ F](id_A))} \text{C}(B, [G \circ F](A)) \\
\varphi_{A,F(A)}(id_A) & \xrightarrow{\varphi_{B,F(B)}} \text{C}(B, [G \circ F](B))
\end{align*}
\] (420)

Since \(\varphi_{A,F(B)}\) is a morphism in \(\text{Set}\), it is a function, and we can apply it in order to define the family

\[
\eta : \text{id}_C \Rightarrow G \circ F, \\
\eta_A := \varphi_{A,F(A)}(id_{F(A)}).
\]

We must show that \(\eta\) is a natural transformation. On the diagram (420), we can start in the top left corner with \(F(id_A)\) and top right corner with \(F(id_B)\) and reach the middle.

We obtain that,

\[
\text{C}(f, [G \circ F](id_A))\left(\varphi_{A,F(A)}(id_A)\right) = \eta_A \circ f
\]

and

\[
\text{C}(id_B, [G \circ F](f))\left(\varphi_{B,F(B)}(id_B)\right) = [G \circ F](f) \circ \eta_B
\]

are equal. That is, the following diagram commutes:

\[
\begin{array}{ccc}
B & \xrightarrow{f} & A \\
\eta_B & & \eta_A \\
\end{array}
\]

\[
[G \circ F](B) \xrightarrow{[G \circ F](f)} [G \circ F](A)
\] (421)

606
In order to define the natural transformation $\varepsilon : F \circ G \Rightarrow \text{id}_D$, we use the inverse transformation $\varphi^{-1}$. For every morphism $g : X \to Y$ in $D$, we have

$$
\begin{align*}
\begin{array}{c}
\varphi^{-1}_{G(X,Y)} \quad \xymatrix{ D(F \circ G)(X,Y) \ar[r]^(.4){(\varphi^{-1}_{G(X,Y)})g} & D(F \circ G)(Y,Y) } \xymatrix{ C(G(X),C(Y)) \ar[r]_(.4){\varphi^{-1}_{G(X,Y)}} & C(G(Y),C(Y)) }
\end{array}
\end{align*}
$$

Define the family

$$
\varepsilon : F \circ G \Rightarrow \text{id}_D,
\varepsilon_X := \varphi^{-1}_{G(X),X}(\text{id}_G(X)).
$$

We can prove that $\varepsilon$ is a natural transformation analogously to how we proved it for $\eta$, and we will skip the details.

We will now show that the triangle diagram (418) commutes. Consider the morphism $(\eta_A, F(id_A))$ in $C^{\text{op}} \times D$. Applying the functors $D(F(-), -)$ and $D(-, G(-))$ to this morphism and using the naturality of $\varphi$, we obtain

$$
\begin{align*}
\begin{array}{c}
\varphi^{-1}_{A,F(A)} \quad \xymatrix{ D([F \circ G \circ F](A),F(A)) \ar[r]^(.4){\varphi^{-1}_{A,F(A)}} & D(F(A),F(A)) } \xymatrix{ C([G \circ F](A),[G \circ F](A)) \ar[r]_(.4){\varphi^{-1}_{A,F(A)}} & C(A,[G \circ F](A)) }
\end{array}
\end{align*}
$$

Note that $\varepsilon_{F(A)}$ is a member of $D([F \circ G \circ F](A), F(A))$.

Composing the functions in (423) in one direction, we obtain

$$
\begin{align*}
\varphi^{-1}_{A,F(A)} \left( C(\eta_A,F(id_A))(\varphi_{G \circ F}(A)(\varepsilon_{F(A)})) \right) = \\
= \varphi^{-1}_{A,F(A)} \left( (\varphi_{G \circ F}(A)(\varepsilon_{F(A)})) \circ \eta_A \right) = \\
= \varphi^{-1}_{A,F(A)}(\text{id}_{G \circ F}(A) \circ \eta_A) = \\
= \text{id}_{F(A)}.
\end{align*}
$$

Composing the functions in (423) in the other direction, we obtain

$$
D(F(\eta_A),F(id_A))(\varepsilon_{F(A)}) = \varepsilon_{F(A)} \circ F(\eta_A).
$$

Therefore,

$$
\text{id}_{F(A)} = \varepsilon_{F(A)} \circ F(\eta_A).
$$

and thus (418) commutes.

We can similarly prove that (419) commutes.

Therefore, $(F, G, \eta, \varepsilon)$ is a unit-counit adjunction.
**Proof that 1189 (b) implies 1189 (a).** Let \((F, G, \eta, \varepsilon)\) be a unit-counit adjunction.

For every pair of objects \(A \in \mathcal{C}\) and \(X \in \mathcal{D}\), define the functions

\[
\varphi_{A,X} : \mathcal{D}(F(A), X) \to \mathcal{C}(A, G(X))
\]

and

\[
\psi^{-1}_{A,X} : \mathcal{C}(A, G(X)) \to \mathcal{D}(F(A), X)
\]

From the naturality of \(\varepsilon\) and from (418), it follows that the following diagram commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{F(\eta_A)} & [F \circ G \circ F](A) \xrightarrow{[F \circ G](\varepsilon)} [F \circ G](X) \\
\downarrow{F(\varphi_{A,X}(g))} & & \downarrow{F(\varphi_{A,X}(g))} \\
F(A) & \xrightarrow{\varepsilon_X} & X \\
\end{array}
\]

Therefore,

\[
g = \varepsilon_X \circ [F \circ G](f) \circ F(\eta_A) = [\varphi_{A,X} \circ \varphi_{A,X}](g)
\]

and thus \(\psi_{A,X}\) is a left inverse of \(\varphi_{A,X}\).

We can analogously show that \(\psi_{A,X}\) is a right inverse, and hence that \(\varphi_{A,X}\) is invertible.

Proposition 1190. The functor \(F : \mathcal{C} \to \mathcal{D}\) is left adjoint to \(G : \mathcal{D} \to \mathcal{C}\) if and only if the dual functor \(F^{op}\) is right adjoint to \(G^{op}\).

This is part of the duality principles listed in proposition 1124.
Proof.

\[ \mathsf{C} \circ (\mathsf{G} \circ (X), A) = \mathsf{C}(A, G(X)) \cong \mathsf{D}(\mathcal{F}(A), X) = \mathsf{D}_{\circ}(X, \mathcal{F}_{\circ}(A)). \]

\[ \square \]

**Definition 1191.** A **concrete category** is a pair \((\mathsf{C}, U)\), where \(\mathsf{C}\) is a category and \(U : \mathsf{C} \to \text{Set}\) is a **faithful functor** that gives us a set for any object of \(\mathsf{C}\). More generally, a pair \((\mathsf{C}, U)\), where \(U : \mathsf{C} \to \mathsf{D}\), is a concrete category over \(\mathsf{D}\).

In the case of a concrete category over \(\text{Set}\), we can regard any morphism \(f\) in \(\mathsf{C}\) as a function. Thus, for example, we can say that \(f\) is an **injective** morphism if \(U(f)\) is an injective function. This is discussed further in proposition 1199.

In the context of a concrete category, we call \(U\) a **forgetful functor** and any left adjoint to \(U\) functor a **free functor**. According to Jean-Pierre Marquis in [Mar19], the motivation for this terminology is that free functors build objects that are free from additional restrictions.

We list several examples in example 1192. The forgetful functor is usually clear from the context, and we identify a concrete category \((\mathsf{C}, U)\) with its underlying set \(\mathsf{C}\). The corresponding free functor, however, often requires a nontrivial but straightforward construction.

**Example 1192.** We list some examples of category adjunctions. Note that only some of them are commonly referred to as “free”.

(a) Perhaps the simplest meaningful example of an adjunction is the **discrete topology** functor \(D : \text{Set} \to \text{Top}\), which is left adjoint to the forgetful functor \(U : \text{Top} \to \text{Set}\), which maps a small topological space \((X, \mathcal{T})\) into its underlying set \(X\).

Given a set \(A\) and a topological space \((X, \mathcal{T})\), every function \(s : A \to X\) is **continuous** when \(A\) is endowed with the discrete topology. Conversely, every continuous function is obviously a **function**. It follows that there is an equality

\[ \text{Top}\left(\underbrace{(A, \text{pow}(A))}_{D(A)}, (X, \mathcal{T})\right) = \text{Set}\left(\underbrace{A, X}_{D(A)}\right). \]

Therefore, \((D, U, \text{id})\) is a hom-adjunction. Furthermore, \((D, U, \text{id}, \text{id})\) is a unit-counit adjunction.

(b) The **indiscrete topology** functor \(I : \text{Set} \to \text{Top}\) is right-adjoint to the same forgetful functor \(U : \text{Top} \to \text{Set}\), again with identities for all natural transformations of the adjunction.

Therefore, we have

\[ D \dashv U \dashv I. \]

(c) We discussed in example 1141 the **discrete category** functor \(D : \text{Set} \to \text{Cat}\). We showed in example 1168 that, when restricted to the subcategory \(\text{DiscrCat}\) rather than \(\text{Cat}\), \(D\) it is an inverse to the forgetful functor \(U\). In the general case, however, this is an adjunction rather than an isomorphism. More precisely, \(D\) is left adjoint to \(U\).
Note that for any functor $F : \mathsf{C} \to \mathsf{D}$, we have $U(F) := F_{|\text{obj}(\mathsf{C})}$. Thus, $U$ is not only a functor in $[\mathsf{Cat}, \mathsf{Set}]$; it also induces a natural isomorphism between the functors $\mathsf{Cat}(D(-), -)$ and $\mathsf{Set}(-, U(-))$.

Indeed, fix a small category $\mathsf{C}$ and a set $A$. From our discussion in example 1168 it is obvious that the restriction $U : \mathsf{Cat}(D(-), \mathsf{C}) \to \mathsf{Set}(A, U(\mathsf{C}))$ is a bijective function.

In order to verify the naturality of the transformation induced by $U$, we must show that for any function $f : B \to A$ and functor $F : \mathsf{C} \to \mathsf{D}$, the following diagram commutes

$$
\begin{array}{ccc}
\mathsf{Cat}(D(A), \mathsf{C}) & \xrightarrow{\mathsf{Cat}(D(f), F)} & \mathsf{Cat}(D(B), \mathsf{D}) \\
\downarrow U & & \downarrow U \\
\mathsf{Set}(A, U(\mathsf{C})) & \xrightarrow{\mathsf{Set}(f, U(F))} & \mathsf{Set}(B, U(\mathsf{D}))
\end{array}
$$

The commutativity of (427) follows from the following: for every functor $S : D(A) \to \mathsf{C}$ we have

$$U(F \circ S \circ D(f)) = U(F) \circ U(S) \circ U(D(f)) = U(F) \circ U(S) \circ f.$$ 

Therefore, $(D, U, U)$ is a hom-adjunction.

We can also explicitly define a unit-counit adjunction. The unit $\eta : \text{id}_{\mathsf{Set}} \Rightarrow U \circ D$ is simply the identity.

The counit is slightly more interesting. Given a small category $\mathsf{C}$, applying $D \circ U$ gives us the subcategory consisting only of the objects and identity morphisms of $\mathsf{C}$. Then the counit $\varepsilon : D \circ U \Rightarrow \text{id}_{\mathsf{Cat}}$ is simply the inclusion functor $I$ from this subcategory to $\mathsf{C}$.

The triangle

$$
\begin{array}{ccc}
D(A) & \xrightarrow{D(\text{id}_A)} & D(A) \\
\downarrow D \circ U \circ D & & \downarrow 1_{D(A)} \\
[D \circ U \circ D](A) & \xrightarrow{1_{[D \circ U \circ D](A)}} & [D \circ U \circ D](A)
\end{array} \quad \quad 
\begin{array}{ccc}
U(\mathsf{C}) & \xrightarrow{U(\text{id}_\mathsf{C})} & U(\mathsf{C}) \\
\downarrow U(\text{id}_\mathsf{C}) & & \downarrow U(\text{id}_\mathsf{C}) \\
[U \circ D \circ U](\mathsf{C}) & \xrightarrow{U(\text{id}_\mathsf{C})} & U(\mathsf{C})
\end{array}
$$

(428)

-corresponding to (418) and (419), obviously commute.

The quadruple $(D, U, \eta, \varepsilon)$ is a unit-counit adjunction.

(d) The left adjoint of the forgetful functor $U : \mathsf{Cat} \to \mathsf{Quiv}$ is the free category functor defined in definition 1337. We denote this functor by $F : \mathsf{Quiv} \to \mathsf{Cat}$.
We can define the family of functions

\[
\varphi : \text{Cat}(F(-), -) \Rightarrow \text{Quiv}(-, U(-)),
\]

\[
\varphi_{Q,C}(S) := \left( v \mapsto S(v), a \mapsto S(t(a)) \right).
\]

(429)

For every functor \( S : F(Q) \to C \), \( \varphi_{Q,C} \) defines a quiver homomorphism that restricts \( S \) to quiver paths of containing only one arc. Formally, \( \iota \) is the canonical embedding

\[
\iota : Q \to [U \circ F](Q)
\]

\[
\iota_V(v) := v,
\]

\[
\iota_A(a) := (h(a), a).
\]

We will later see that \( \iota \) is the unit of a unit-counit adjunction.

Now, from (564), it is clear that the free category functor \( F \), when restricted to the set of quiver homomorphisms \( \text{Quiv}(Q, U(C)) \), is the two-sided inverse of \( \varphi_{Q,C} \).

We will show that \( \varphi \) is a natural transformation. Fix a functor \( G : C \to D \) and a homomorphism \( (g_V, g_A) : Q \to R \). We must show that the following diagram commutes:

\[
\begin{array}{ccc}
\text{Cat}(F(Q), C) & \xrightarrow{\varphi_{Q,C}} & \text{Cat}(F(R), D) \\
\text{Set}(Q, U(C)) & \xrightarrow{\text{Set}(g_V, g_A, U(G))} & \text{Set}(R, U(D))
\end{array}
\]

(430)

That is, for every functor \( S : F(A) \to C \), we must show

\[
\varphi_{R,D}(G \circ S \circ F(g_V, g_A)) = U(G) \circ \varphi_{R,D}(S) \circ (g_V, g_A).
\]

This is also clear from (564).

Therefore, \( (F, U, \varphi) \) is a hom-adjunction.

Furthermore, the canonical embedding \( \iota \) defined above, when parameterized by \( Q \), is a unit of adjunction.

The counit \( \varepsilon : F \circ U \Rightarrow \text{id}_{\text{Cat}} \) is more involved. As discussed in definition 1337, for every finite path \( p \) in the quiver \( U(C) \) with arcs \( a_1, ..., a_n \), the functor \( F \circ U \) simply “evaluates” \( p \) as

\[
a_n \circ a_{n-1} \circ \cdots \circ a_1.
\]

Since the embedding only produces paths with a single arc, the adjunction triangles commute:

\[
\begin{array}{ccc}
F(Q) & \xrightarrow{F(\iota_Q)} & F(Q) \\
\downarrow F(\iota_Q) & & \downarrow F(\iota_Q) \\
[U \circ F \circ U](Q) & \xrightarrow{\varepsilon_{F(Q)}} & [U \circ F \circ U](Q)
\end{array}
\]

\[
\begin{array}{ccc}
U(C) & \xrightarrow{U(\iota_C)} & U(C) \\
\downarrow U(\iota_C) & & \downarrow U(\iota_C) \\
[U \circ F \circ U](C) & \xrightarrow{\varepsilon_{U(C)}} & [U \circ F \circ U](C)
\end{array}
\]

(431)
(e) In definition 1320 (f), we have defined the forgetful functor $U : \text{Quiv} \rightarrow \text{MultGph}$. Given a Grothendieck universe $U$ and a choice function $c$ for the family of two-element sets in $U$, we have an orientation functor $O_c : \text{MultGph} \rightarrow \text{Quiv}$ defined in definition 1323.

It may seem that, for a fixed choice function, $O_c$ is left adjoint to $U$. This is not true, however, as shown in fig. 30.

**Definition 1193.** We call the quadruple $(F, G, \eta, \varepsilon)$ with signature (402) an **adjoint equivalence** if it is both an adjunction and equivalence.

**Proposition 1194.** Let $(F, G, \eta, \varepsilon)$ be a category equivalence between $\mathbf{C}$ and $\mathbf{D}$.

There exists a natural isomorphism $\zeta : \text{id}_\mathbf{C} \Rightarrow G \circ F$ such that $(F, G, \zeta, \varepsilon)$ is an adjoint equivalence.

**Proof.** From proposition 1173 it follows that $F$ is fully faithful and essentially surjective.

We will now use the same trick as in the end of the proof of theorem 1175 (Fully faithful and essentially surjective functor induces equivalence) to define $\zeta$.

Since $F$ is fully faithful, there is a bijective function

$$\varphi : D(F(A), [F \circ G \circ F](A)) \rightarrow C(A, [F \circ G](A)).$$

Hence, we can define

$$\zeta : \text{id}_\mathbf{C} \rightarrow G \circ F,$$

$$\zeta_A := \varphi(\varepsilon^{-1}_{F(A)})$$

so that $F(\zeta_A) = \varepsilon^{-1}_{F(A)}$. By proposition 1152 (h), $\zeta_A$ is also an isomorphism.

As in theorem 1175 (Fully faithful and essentially surjective functor induces equivalence), we use proposition 1151 and the naturality of $\varepsilon$ to prove that (412) implies (411) (with $\eta$ replaced by $\zeta$).

Therefore, $\zeta$ is a natural isomorphism and the quadruple $(F, G, \zeta, \varepsilon)$ is an equivalence of categories.

**Proposition 1195.** If a functor has two left adjoints (resp. right adjoints), then there exists a unique natural isomorphism between them.

We say that left adjoints (resp. right adjoints) are unique up to a unique natural isomorphism.
Proof: We will first prove the statement for left adjoints. Suppose that \((F', G, \eta', \varepsilon')\) and \((F'', G, \eta'', \varepsilon'')\) are two unit-counit adjunctions.

**Proof of existence of isomorphism.** We can utilize the naturality of the units \(\eta'\) and \(\eta''\) and counits \(\varepsilon'\) and \(\varepsilon''\) to show that the following diagram commutes:

By the commuting triangle (418), all paths from \(F'(A)\) to \(F''(A)\) above are identities. The bottom-most path in (432) justifies defining the natural transformation

\[
\alpha : F' \Rightarrow F'' \\
\alpha_A := \varepsilon''_{F''(A)} \circ F'(\eta''_A).
\]

Then \(\alpha_A\) is an isomorphism for every object \(A\) in \(\mathbf{C}\) with inverse \(\varepsilon''_{F''(A)} \circ F'(\eta''_A)\). Therefore, it is a natural isomorphism from \(F'\) to \(F''\).

**Proof of uniqueness of isomorphism.** Suppose that \(\beta : F' \Rightarrow F''\) is another natural isomorphism. Then, by the commuting triangle (418), the following diagram also commutes:

Therefore,

\[
\beta_A = \varepsilon''_{F''(A)} \circ F''(\eta''_A) \circ \alpha_A = (418) \alpha_A.
\]

This finishes the proof for left adjoints. The other direction is dual. If \(G'\) and \(G''\) are two right adjoints to \(F\), then by proposition 1190, \(G'\text{op}\) and \(G''\text{op}\) are left adjoints and are thus isomorphic. Then by proposition 1127, \(G'\) and \(G''\) are also isomorphic. \(\square\)
**Proposition 1196.** Fix a category \( C \). We can characterize the universal objects in \( C \) from definition 1130 via adjunctions with the terminal category \( 1 \).

Let \( \Delta^1 : C \to 1 \) be the constant functor into \( 1 \).

(a) The object \( I \) of \( C \) is initial if and only if it is (the unique value of) a left adjoint to \( \Delta^1 \) functor.

In particular, the uniqueness proved in proposition 1133 (a) follows from proposition 1195.

(b) Dually, the terminal objects are exactly the right adjoint to \( \Delta^1 \) functors.

**Proof.** We will only prove proposition 1196 (a) since the other direction is dual.

**Proof of sufficiency.** Let \( I \) be an initial object in \( C \). We can then regard it as a functor \( F : 1 \to C \).

We define the natural transformations

\[
\eta : \text{id}_1 \Rightarrow \Delta^1_I \circ F \\
\eta_0 := \text{id}_0
\]

and

\[
\varepsilon : F \circ \Delta^1_I \Rightarrow \text{id}_C \\
\varepsilon_A \text{ is the unique morphism } I \to A
\]

Since \( I \) has a unique morphism into any other object of \( C \), for every morphism \( f : A \to B \), the following diagram commutes:

\[
\begin{array}{c}
I \\
\varepsilon_A \downarrow \\
A \\
f \downarrow \\
B
\end{array}
\xymatrix{
I \ar[r]^-{\text{id}_I} & I \\
A \ar[r]_f & B
}
\]

(434)

It follows that both \( \eta \) and \( \varepsilon \) are natural transformations. Furthermore, they trivially satisfy the triangle diagrams triangle (418) and (419).

Hence, \( (F, \Delta^1_I, \eta, \varepsilon) \) is a unit-counit adjunction.

**Proof of necessity.** Conversely, suppose that \( (F, \Delta^1_I, \eta, \varepsilon) \) is a unit-counit adjunction.

Let \( I := F(0) \). Then \( \varepsilon_A \) is a morphism from \( I \) to \( A \). By (418), \( \varepsilon_I = \text{id}_I \) since the following diagram commutes:

\[
\begin{array}{c}
I \\
\varepsilon_{\Delta^1(0)} \downarrow \\
\Delta^1(0) \\
\varepsilon_{\Delta^1(\eta_0)} \downarrow \\
\Delta^1(\eta_0)
\end{array}
\xymatrix{
I \ar[r]^-{\text{id}_I} & I \\
\Delta^1(0) \ar[r]_-{\text{id}_{\Delta^1(0)}} & \Delta^1(0)
}
\]

(435)
Suppose that $\zeta$ is another morphism from $I$ to $A$. The naturality of $\varepsilon$ implies that, for the morphism $\text{id}_A : A \to A$, the following diagram commutes:

$$
\begin{array}{ccc}
I & \xrightarrow{\text{id}} & I \\
\downarrow{\varepsilon_A} & & \downarrow{\varepsilon_A} \\
I & \xrightarrow{\zeta} & A
\end{array}
$$

(436)

The upper left triangle in (436) is (435).

We conclude that $\zeta = \varepsilon_A$ and, generalizing on $A$, that every morphism from $I$ is unique. \hfill \square

Remark 1197. We discussed in example 1131 (b) that “the” trivial group $\{e\}$ is a zero object of $\text{Grp}$. By proposition 1196, this object induces a functor that is both left adjoint and right adjoint of $\Delta_1$. Nevertheless, the categories $\text{Grp}$ and 1 are not equivalent.

Remark 1198. We will now regard adjoint functors as a way to “construct” new objects.

Let $(F, G, \iota, \pi)$ be a unit-counit adjunction between the categories $\mathcal{C}$ and $\mathcal{D}$. In the current context, especially in connection with limits and colimits, we will call the components of the counit $\pi : F \circ G \Rightarrow \text{id}_D$ — projections, and the components of the unit $\iota : \text{id}_C \Rightarrow G \circ F$ — coprojections.

Take objects $A$ in $\mathcal{C}$ and $X$ in $\mathcal{D}$ and a morphism $f : A \to G(X)$. We want to obtain a morphism $\widetilde{f} : F(A) \to X$, for which the following diagram commutes:

$$
\begin{array}{ccc}
A & \xrightarrow{f} & G(X) \\
\downarrow{\iota_A} & & \downarrow{G(\widetilde{f})} \\
[G \circ F](A) & & [G \circ F](A)
\end{array}
$$

(437)

From the naturality of $\iota$ and from the triangle diagram (419) it follows that the following diagram commutes:

$$
\begin{array}{ccc}
A & \xrightarrow{f} & G(X) & \xrightarrow{\text{id}_{G(X)}} & G(X) \\
\downarrow{\iota_A} & & \downarrow{G(\widetilde{f})} & & \downarrow{G(\pi_X)} \\
[G \circ F](A) & \xrightarrow{G(\iota_A)} & [G \circ F \circ G](X) & \xrightarrow{[G \circ F](f)} & [G \circ F \circ G](X)
\end{array}
$$

(438)

It is clear from (438) that

$$G(\widetilde{f}) = G(\pi_X) \circ [G \circ F](f) = G(\pi_X \circ F(f)).$$

Furthermore, this value is unique. From the naturality of $\pi$ and the triangle diagram (418)
it follows that the following diagram commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{id_{F(A)}} & F(A) \\
\downarrow{\iota_{F(A)}} & \downarrow{F(f)} & \downarrow{\pi_{F(A)}} \\
[F \circ G \circ F](A) & \xrightarrow{\pi_{F(A)}} & F(X) \\
\downarrow{[F \circ G \circ f]} & \downarrow{\pi_X} & \downarrow{X} \\
[F \circ G](X) & \xrightarrow{\pi_X} & X
\end{array}
\]

Therefore,

\[\tilde{f} = \pi_X \circ F(f)\]

Taking into account that the functor \(F\) itself is unique up to a unique isomorphism, as per proposition 1195, we have proved the following statement:

For every object \(A\) in \(\mathcal{C}\), there exist unique up to a unique isomorphism object \(F(A)\) in \(\mathcal{D}\) and canonical coprojection map \(\iota_A : A \to [G \circ F](A)\) satisfying the following property, called a **universal mapping property**:

For every object \(X\) in \(\mathcal{D}\) and every map \(f : A \to G(X)\) in \(\mathcal{C}\), there exists a unique map \(\tilde{f} : F(A) \to X\) in \(\mathcal{D}\) such that the diagram (437) commutes.

Intuitively, this universal mapping property states that any map (morphism) with domain \(A\) in \(\mathcal{C}\) can be transformed into a map with domain \(F(A)\) in \(\mathcal{D}\).

The statement becomes more meaningful when we regard \(G : \mathcal{D} \to \mathcal{C}\) as a forgetful functor. In this case, every object of \(\mathcal{D}\) is regarded as an object of \(\mathcal{C}\), and we write \(X\) rather than \(G(X)\). The universal mapping property then becomes:

For every object \(A\) in \(\mathcal{C}\), there exist unique up to a unique isomorphism object \(F(A)\) in \(\mathcal{D}\) and canonical coprojection map \(\iota_A : A \to F(A)\) satisfying the following universal mapping property:

For every object \(X\) in \(\mathcal{D}\) and every map \(f : A \to X\) in \(\mathcal{C}\), there exists a unique map \(\tilde{f} : F(A) \to X\) in \(\mathcal{D}\) such that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{f} & X \\
\downarrow{\iota_A} & \downarrow{\tilde{f}} & \downarrow{f} \\
F(A) & & \end{array}
\]

(440)

In definition 1191 we mentioned that we will call the left adjoint of a forgetful functor a free functor. Universal mapping properties allow characterizing certain “free constructions”, such as the free groups defined in definition 476, without explicitly building a free functor and proving that it is left adjoint. Indeed, for every suitable object and map, we explicitly build the natural isomorphism \(\varphi\) of a hom-adjunction, and the commutative triangle (437) ensures that this \(\varphi\) is a natural transformation.
Universal mapping properties of this form are used for colimits — see remark 1207.
Of course, there is a dual universal mapping property:

For every object $X$ in $\mathbb{D}$ there exist unique up to a unique isomorphism object
$G(X)$ in $\mathbb{C}$ and canonical projection map $\pi_X : [F \circ G](X) \to X$ satisfying the
following property, called a **universal mapping property**:

For every object $A$ in $\mathbb{C}$ and every map $g : F(A) \to X$ in $\mathbb{D}$, there exists
a unique map $\tilde{g} : A \to G(X)$ in $\mathbb{C}$ such that the following diagram
commutes:

$$
\begin{array}{ccc}
F(A) & \xrightarrow{\tilde{g}} & G(X) \\
\downarrow_{F(\tilde{g})} & & \downarrow_{\pi_X} \\
[F \circ G](X) & \xrightarrow{g} & X
\end{array}
$$

In this case, we can regard $F$ as a forgetful functor and $G$ as a free functor to obtain the
following:

For every object $X$ in $\mathbb{D}$ there exist unique up to a unique isomorphism object
$G(X)$ in $\mathbb{C}$ and canonical projection map $\pi_X : G(X) \to X$ satisfying the following
property, called a **universal mapping property**:

For every object $A$ in $\mathbb{C}$ and every map $g : A \to X$ in $\mathbb{D}$, there exists
a unique map $\tilde{g} : A \to G(X)$ in $\mathbb{C}$ such that the following diagram
commutes:

$$
\begin{array}{ccc}
a & \xrightarrow{\tilde{g}} & G(X) \\
\downarrow_{g} & & \downarrow_{\pi_X} \\
A & \xrightarrow{\pi_X} & X
\end{array}
$$

Universal mapping properties of this form are used for limits — see remark 1207.

**Proposition 1199.** If $\mathbb{C}$ is a concrete category with forgetful functor $U : \mathbb{C} \to \text{Set}$.

(a) Every **nonempty injective morphism** in $\mathbb{C}$ is a **monomorphism**.

That is, for every morphism $f$ in $\mathbb{C}$, if $U(f)$ is a **nonempty injective function**, $f$ is a
**monomorphism**.

(b) Dually, every **surjective morphism** in $\mathbb{C}$ is an **epimorphism**.

(c) Every **split monomorphism** is injective, every **split epimorphism** is surjective and every
categorical isomorphism is a **bijective function**.

(d) If $U$ has a left adjoint, every monomorphism is injective.

(e) If $U$ has a right adjoint, every epimorphism is surjective.
Proof.

**Proof of 1199 (a).** By proposition 989, a nonempty injective function is a monomorphism in \( \textbf{Set} \). This translates to morphisms in \( \mathbf{C} \) since they are special cases of functions. Formally, this follows from proposition 1152 (f).

**Proof of 1199 (b).** Again by proposition 989, a surjective function is an epimorphism in \( \textbf{Set} \). Again from proposition 1152 (f), this implies that every surjective morphism in \( \mathbf{C} \) is an epimorphism.

**Proof of ???.** Follows from proposition 1152 (d).

**Proof of 1199 (d).** Let \( F : \textbf{Set} \to \mathbf{C} \) be a left adjoint to \( U \) and fix a set \( A \). As discussed in remark 1198, this leads to the following universal mapping property:

For every object \( X \) in \( \mathbf{D} \) and every morphism \( g : A \to U(X) \) in \( \mathbf{C} \), there exists a unique morphism \( \tilde{g} : F(A) \to X \) such that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{g} & U(X) \\
\downarrow{\iota_A} & & \downarrow{U(\tilde{g})} \\
[U \circ F](A) & \xleftarrow{U(\tilde{id}_X)} & U(\tilde{g})
\end{array}
\]  

(443)

First, suppose that \( f : X \to Y \) is a monomorphism in \( \mathbf{C} \) and let \([U(f)](x_1) = [U(f)](x_2)\) (by proposition 989 (a), \( U(f) \) is nonempty). Aiming at a contradiction, suppose that \( x_1 \neq x_2 \). Define the function \( g : U(X) \to U(X) \) as the identity \( U(\text{id}_X) \) for which \( x_1 \) and \( x_2 \) are swapped as in the proof of proposition 989 (c). By the property (443), the following diagram commutes:

\[
\begin{array}{ccc}
[U \circ F \circ U](X) & \xleftarrow{U(\tilde{id}_X)} & U(X) \\
\downarrow{U(\tilde{id}_X)} & & \downarrow{U(\tilde{id}_X)} \\
U(X) & \xrightarrow{U(\tilde{g})} & U(X) \\
\downarrow{U(\tilde{id}_X)} & & \downarrow{U(\tilde{g})} \\
[U \circ F \circ U](X) & \xleftarrow{U(\tilde{g})} & U(Y) \\
\end{array}
\]

(444)

In particular,

\[ U(f) \circ U(\text{id}_X) = U(f) \circ U(\tilde{g}) , \]

and thus \( U(\text{id}_X) = U(\tilde{g}) \) and \( U(\text{id}_X) = g \). But this contradicts our assumption that \( g \) differs from the identity by construction. Therefore, \( U(f) \) must be an injective function.
**Proof of 1199 (e).** Dually, let $F$ be a right adjoint to $U$. It follows from proposition 1190, proposition 1124 (a) and proposition 1199 (d) that every epimorphism is a surjection.

We will also give a direct proof. Since $F$ is right adjoint, we have the following universal mapping property:

For every object $X$ in $\mathbf{C}$ and every function $h : A \to U(X)$ in $\mathbf{Set}$, there exists a unique morphism $\tilde{h} : F(A) \to X$ in $\mathbf{C}$ such that the following diagram commutes:

$$
\begin{array}{ccc}
U(X) & \xrightarrow{\pi_X} & A \\
\uparrow{U(h)} & & \\
[U \circ F](A) & \xrightarrow{h} & X \\
\end{array}
$$

Suppose that $f$ is an epimorphism and that there exists some value $y_0 \in U(Y)$ not in the image of $f$. As in the proof of proposition 989 (e), fix some set $z$ not in $U(Y)$, define $S := U(Y) \cup \{z\}$ and define $g : U(Y) \to S$ by swapping $y_0$ and $z$ in the identity $\text{id}_Z$. Let $s : U(Y) \to S$ be the restriction of $\text{id}_Z$ to $U(Y)$.

By the property (445), the following diagram commutes:

$$
\begin{array}{ccc}
U(X) & \xrightarrow{U(f)} & U(Y) \\
\uparrow{s} & & \downarrow{h} \\
U(Y) & \xrightarrow{\text{id}_Z} & S \\
\uparrow{U(\tilde{h})} & & \downarrow{\pi_S} \\
[U \circ F](S) & \xrightarrow{\pi_S} & [U \circ F](S) \\
\end{array}
$$

In particular,

$$
U(\text{id}_Z) \circ U(f) = U(\tilde{h}) \circ U(f),
$$

and thus $U(\text{id}_Z) = U(\tilde{h})$ and $\text{id}_Z = h$. But this contradicts our construction of $h$. Therefore, $U(f)$ must be a surjective function.
14.5. Categorical limits

Definition 1200. Comma categories allow us to define morphisms between morphisms, which becomes useful in, for example, the concise definition of a limit in definition 1202 (c).

(a) We first prove the most general construction. Let \( F : \mathcal{A} \to \mathcal{C} \) and \( G : \mathcal{B} \to \mathcal{C} \) be any two functors with a common codomain. We define their comma category \((F \downarrow G)\) as follows:

- The set of objects \( \text{obj}(F \downarrow G) \) is the set of triples \((A, s, B)\), where \( A \in \mathcal{A}, B \in \mathcal{B} \) and \( s : F(A) \to G(B) \).
- The set of morphisms \([F \downarrow G]\)((A, s, B), (A', s', B')) is the set of pairs \((f : A \to A', g : B \to B')\), such that the following diagram commutes:

\[
\begin{array}{ccc}
F(A) & \xrightarrow{f} & F(A') \\
\downarrow s & & \downarrow s' \\
G(B) & \xrightarrow{g} & G(B')
\end{array}
\]

(b) It is often the case where either \( F \) or \( G \) are constant. If, instead of \( G \), we are given an object \( X \) in \( \mathcal{C} \), we use the constant functor \( \Delta_X : 1 \to \mathcal{C} \) with domain the terminal category \( 1 \), in order to use the comma category \((F \downarrow \Delta_X)\).

We can thus simplify definition 1200 (a) as follows:

\[
\text{obj}(F \downarrow X) := (F \downarrow \Delta_X), \quad \text{obj}(X \downarrow G) := (\Delta_X \downarrow G)
\]

\[
[F \downarrow X]((A, s), (A', s')) \quad \text{is the set of } f : A \to A', \text{ such that }
\]

\[
X \xleftarrow{s} \xrightarrow{F(f)} F(A) \xleftarrow{s'} \quad (448)
\]

\[
[X \downarrow G]((s, B), (s', B')) \quad \text{is the set of } g : B \to B', \text{ such that }
\]

\[
G(B) \xleftarrow{G(g)} \xrightarrow{s'} \xleftarrow{s} X \quad (449)
\]
Definition 1201. Consider a morphism \( f : A \to B \). We say that \( f \) factors through the object \( X \) if there exist morphisms \( g : A \to X \) and \( h : X \to B \) such that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{g} & X \\
\downarrow{f} \quad & & \downarrow{h} \\
B
\end{array}
\]  

(450)

If, given \( g, h \) is unique or vice versa, we say that \( f \) uniquely factors through \( X \). In some cases, for example in definition 454 (a), \( f \) there exists unique \( g \) and \( h \) and neither of them must be given beforehand.

Definition 1202. Let \( D : I \to C \) be a diagram.

(a) We define category of cones to \( D \) as the constant-functor comma category

\[
\text{Cone}(D) := (\Delta^I \downarrow D),
\]

where \( \Delta^I : C \to [I, C] \) is the I-shaped diagonal functor on \( C \).

(b) A \( D \)-cone is simply a member of \( \text{Cone}(D) \).

Explicitly, a cone with vertex \( A \in C \) is a pair \( (\Delta_I^A, \alpha) \), where \( \Delta^A_I \) is the constant functor at \( A \) and \( \alpha \) is a natural transformation from \( \Delta^A_I \) to \( D \).

Even more explicitly, a cone is a family of morphisms

\[
\{ \alpha_k : A \to D(k) \}_{k \in I}.
\]

(451)

satisfying a simplified naturality condition (compared to (381)). For every morphism \( u : k \to m \) in \( I \), the following diagram must commute:

\[
\begin{array}{ccc}
A & \xrightarrow{\alpha_k} & D(k) \\
\downarrow{\alpha_m} \quad & & \downarrow{D(u)} \\
D(m)
\end{array}
\]  

(452)

Note that this diagram is very different from (448), they merely look similar.

(c) A limit cone of the diagram \( D \) is a terminal object of the cone category \( \text{Cone}(D) \).

Explicitly, \((L, \lambda)\) is limit cone if, for every cone \((A, \alpha)\), there exists a unique cone morphism \( l_A : (A, \alpha) \to (L, \lambda) \).

Even more explicitly, \((L, \lambda)\) is a limit cone if it satisfies the following universal mapping property:
For every cone \((A, \alpha)\), there exists a unique morphism \(l_A : A \to L\) such that following diagram commutes for every index morphism \(u : k \to m\):

\[
\begin{array}{c}
A \\
\downarrow l_A \\
L \\
\downarrow \lambda_k \\
D(k)
\end{array}
\quad \middle\rightarrow
\begin{array}{c}
\downarrow \lambda_m \\
D(m)
\end{array}
\]

\[
\begin{array}{c}
\alpha_k \quad \alpha_m
\end{array}
\]

Thus, \(\alpha_k\) \textit{uniquely factors through} \(L\) for every \(k \in I\). Furthermore, \(l_A\) is compatible with morphisms in \(I\) and does not depend on \(k\).

From proposition 1133 (b) and proposition 1196 (b) it follows that a limit, if it exists, is unique up to a unique isomorphism.

Without further context, we usually refer to \(L\) as the limit vertex and \((L, \lambda)\) as the limit cone. By “the limit”, we usually mean the vertex \(L\).

(d) \textbf{Dually}, a \textit{D-cocone} with vertex \(A\) is a family of morphisms

\[
\{\alpha_k : D(k) \to A\}_{k \in I}.
\]

satisfying the naturality condition that for every morphism \(u : k \to m\) in \(I\), the following diagram must commute:

\[
\begin{array}{c}
D(k) \\
\downarrow D(u) \\
A \\
\downarrow \alpha_k \\
D(m)
\end{array}
\quad \middle\rightarrow
\begin{array}{c}
\downarrow \alpha_m \\
D(m)
\end{array}
\]

\[
\begin{array}{c}
\alpha_k \\
\downarrow l_A \\
L
\end{array}
\quad \middle\rightarrow
\begin{array}{c}
\downarrow \lambda_m \\
D(m)
\end{array}
\]

\[
\begin{array}{c}
\alpha_k \quad \alpha_m
\end{array}
\]

(e) \textbf{A colimit cocone} of the diagram \(D\) is an \textit{initial object} of the cocone category \((D \downarrow \Delta)\).

There are two major differences compared to limits: a colimit cocone is an initial object, not a terminal object, and its underlying comma category is \((D \downarrow \Delta)\), not \((\Delta^I \downarrow D)\).

The analogous diagram to (453) is exactly its \textbf{opposite}:

\[
\begin{array}{c}
D(k) \\
\downarrow l_A \\
L
\end{array}
\quad \middle\rightarrow
\begin{array}{c}
\downarrow \lambda_k \\
D(k)
\end{array}
\]

\[
\begin{array}{c}
\alpha_k \quad \alpha_m
\end{array}
\]

\[
\begin{array}{c}
\downarrow \lambda_m \\
D(m)
\end{array}
\]

In particular, \(\alpha_k\) \textit{uniquely factors through} \(L\) for every \(k \in I\).

Without further context, we usually refer to \(L\) as the colimit vertex and \((L, \lambda)\) as the colimit cocone. By “the colimit”, we usually mean the vertex \(L\).
**Proposition 1203.** For every cone \((A, \alpha)\) of the diagram \(D\) in \(C\), \((A, \alpha^{\text{op}})\) is a cocone of \(D^{\text{op}}\) in the opposite category \(C^{\text{op}}\).

Even more, for every limit \((L, \lambda)\) of \(D\) in \(C\), \((L, \lambda^{\text{op}})\) is a colimit of \(D^{\text{op}}\) in \(C^{\text{op}}\).

This is part of the duality principles listed in proposition 1124.

**Proof.** Note that the defining diagrams (451), (452) and (453) are dual to (454), (455) and (456).

**Lemma 1204.** Any two limits (resp. colimits) of a diagram are isomorphic.

We prove a stronger result in corollary 1206.

**Proof.** Let \((L', \lambda')\) and \((L'', \lambda'')\) be two limit cones over the diagram \(D : I \to C\). The definition (453) of a limit implies that

\[
\begin{commutative diagram}
  L'' & \xleftarrow{\lambda''_k} & L' & \xrightarrow{\lambda'_k} & L'' \\
  \downarrow{\lambda''_k} & & \downarrow{\lambda'_k} & & \downarrow{\lambda''_k} \\
  D(k) & \xrightarrow{D(k)} & D(m) & \xleftarrow{D(k)} & D(m)
\end{commutative diagram}
\]

commutes.

Therefore, the limits \(L'\) and \(L''\) are isomorphic.

Now let \((L', \lambda')\) and \((L'', \lambda'')\) be colimit cocones. By proposition 1203, \((L', \lambda'^{\text{op}})\) and \((L'', \lambda''^{\text{op}})\) are limits in the opposite category and are thus isomorphic. By proposition 1127, the colimits are isomorphic.

**Proposition 1205.** Suppose that, for a given category \(C\), the limits over all \(I\)-shaped diagrams exist. Denote by \(\lim(D)\) the vertex of the limiting cone of the diagram \(D\).

Given a natural transformation \(\alpha : D \to E\) between diagrams, the diagram (459) defining a limit uniquely determines a morphism \(\lim(D) \to \lim(E)\). Denote this morphism by \(\lim(\alpha)\).

We have defined a functor

\[
\lim : [I, C] \to C
\]

This functor is right adjoint to the diagonal functor

\[
\Delta : C \to [I, C]
\]

Dually, the colimit functor

\[
\text{colim} : [I, C] \to C
\]

is left adjoint to \(\Delta\).

**Proof.** It is sufficient to prove this for limits since the statement for colimits follows from the duality principles proposition 1190 and proposition 1203. The unit \(\eta : \text{id}_C \to [\lim \circ \Delta]\) of the adjunction

\[
\Delta \dashv \lim
\]

623
is the unique morphism from an object $A$ of $\mathbf{C}$ to the limit of its constant diagram $\Delta_A : I \to \mathbf{C}$ such that (453) commutes. The counit is more complicated because its components are themselves natural transformations:

$$\varepsilon : [\Delta \circ \lim] \Rightarrow \text{id}_{[I,C]}$$

$$\varepsilon_D : \Delta(\lim D) \Rightarrow D$$

$$\varepsilon_{D,k} : \lim(D) \to D(k),$$

where $\varepsilon_{D,k}$ are the projections of the limit.

For any diagram $D : I \to \mathbf{C}$, the following triangle commutes:

$$\begin{array}{ccc}
\lim(D) & \xrightarrow{\eta_{\lim(D)}} & \lim(D) \\
\downarrow{\eta_{\lim(D)}} & & \downarrow{\eta_{\lim(D)}} \\
[\lim \circ \Delta \circ \lim](D) & & [\lim \circ \Delta \circ \lim](D)
\end{array}$$

Indeed,

$$\eta_{\lim(D)} : \lim(D) \to [\lim \circ \Delta \circ \lim](D)$$

is the unique morphism such that (453) commutes, and (somewhat) similarly for

$$\lim(\varepsilon_D) : [\lim \circ \Delta \circ \lim](D) \to \lim(D).$$

It follows that both $\lim(D)$ and $[\lim \circ \Delta \circ \lim](D)$ are limits over the same diagram. By proposition 1205, they are isomorphic, and hence (458) commutes.

Also, for any object $A$ in $\mathbf{C}$, the following triangle also commutes:

$$\begin{array}{ccc}
\Delta(A) & \xrightarrow{\Delta(\eta_A)} & \Delta(A) \\
\downarrow{\Delta(\eta_A)} & & \downarrow{\Delta(\eta_A)} \\
[\Delta \circ \lim \circ \Delta](A) & & [\Delta \circ \lim \circ \Delta](A)
\end{array}$$

Indeed,

$$\Delta(\eta_A) = \{\eta_A : A \to \lim(\Delta_A)\}_{k \in I}$$

is the constant family consisting of $\eta_A$ and

$$\varepsilon_{\Delta(A)} = \{\varepsilon_{\Delta(A)} : \lim(\Delta_A) \to \Delta_A(k)\}_{k \in I}.$$
Proof. Follows from proposition 1195 and proposition 1205.

Remark 1207. The limit diagram (453) may seem unrelated to the universal mapping properties discussed in remark 1198, however it is actually a special case.

Suppose that, for a given category \( \mathcal{C} \), the limits over all \( \mathbf{I} \)-shaped diagrams exist and fix a diagram \( D : \mathbf{I} \to \mathcal{C} \). Consider the functors

\[
\text{lim} : [\mathbf{I}, \mathcal{C}] \to \mathcal{C} \\
\Delta : \mathcal{C} \to [\mathbf{I}, \mathcal{C}]
\]
discussed in proposition 1205. We have established that \( \Delta \dashv \text{lim} \).

For every diagram \( D : \mathbf{I} \to \mathcal{C} \), there exist unique up to a unique isomorphism object \( \text{lim}(D) \) in \( \mathcal{C} \) and canonical projection map \( \lambda : [\Delta \circ \text{lim}(D)] \to D \) satisfying the following universal mapping property:

For every object \( A \) in \( \mathcal{C} \) and every natural transformation \( \alpha : \Delta(A) \Rightarrow D \), there exists a unique morphism \( l : A \to \text{lim}(D) \) such that the following diagram commutes:

\[
\begin{array}{ccc}
D & \xrightarrow{\alpha} & \Delta(A) \\
\downarrow{} & {} & \downarrow{\lambda} \\
[\Delta \circ \text{lim}(D)] & \xrightarrow{\Delta(l)} & D
\end{array}
\]  \hspace{1cm} (460)

For a fixed index object \( k \in \mathbf{I} \), this becomes:

\[
\begin{array}{ccc}
D(k) & \xrightarrow{\alpha_k} & A \\
\downarrow{\lambda_k} & & \downarrow{l} \\
\text{lim}(D) & \xrightarrow{l_k} & \text{lim}(D)
\end{array}
\]  \hspace{1cm} (461)

The defining diagram (453) of a limit simply encodes the cone naturality condition (452) into (461).

Except for being simpler to check, limits defined via universal mapping properties have the advantage (compared to adjoint functors) that limits can exist for some diagrams and not for others.

The construction for colimits is dual.

Proposition 1208. The limits of an empty diagram over \( \mathcal{C} \) are the terminal objects and the colimits are the initial object.

Compare this result with proposition 1209.

Proof. A cone of the empty diagram is simply an object of \( \mathcal{C} \). A limit cone is then an object \( L \) such that every other object \( A \) has a unique morphism \( l_A : A \to L \). This is precisely the definition of a terminal object.

The statement for colimits follows by duality.
**Proposition 1209.** Fix an arbitrary category $\mathbf{C}$.

(a) If $I$ is an initial object of $\mathbf{C}$, $(I, \xi)$ is a limit cone of the identity functor $\text{id}_\mathbf{C}$, where

$$\xi := \{\xi_A : I \to A\}_{A \in \mathbf{C}}$$

is the family of unique morphisms with domain $I$.

(b) Conversely, if $(L, \lambda)$ is a limit cone of the identity, then $L$ is an initial object.

Dually, by proposition 1132 and proposition 1203, the cocones (and colimits) of the identity functor are the terminal objects.

Compare this result with proposition 1208.

Proof.

**Proof of 1209 (a).** For an initial object $I$ with morphisms $\xi$, and for any morphism $f : B \to C$, from the uniqueness of the arrows in $\xi$ it follows that $\xi_C = f \circ \xi_B$. Thus, the following naturality diagram commutes:

$$\begin{array}{ccc}
B & \xrightarrow{f} & C \\
\xi_B & \searrow & \xi_C \\
\uparrow & \downarrow & \uparrow \\
I & \xrightarrow{\xi_I} & I
\end{array}$$

(462)

Therefore, $(I, \xi)$ is a cone.

Now let $(A, \alpha)$ be another cone. The naturality of $\alpha$ implies that the following diagram commutes for every object $B$:

$$\begin{array}{ccc}
A & \xrightarrow{\alpha_I} & I \\
\sigma_A & \searrow & \downarrow \\
B & \xrightarrow{\xi_B} & I
\end{array}$$

(463)

In particular, since $\xi_I = \text{id}_I$, for every morphism $g : A \to I$ we have

$$\begin{array}{ccc}
A & \xrightarrow{\alpha_I} & I \\
g & \searrow & \downarrow \\
I & \xrightarrow{\text{id}_I} & I
\end{array}$$

(464)

which shows that $\alpha_I = g$ and thus $\alpha_I$ is the unique morphism from $A$ to $I$. 

626
Therefore, for any morphism \( f : B \to C \), the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{\sigma_B} & C \\
\downarrow{\xi_B} & & \downarrow{\xi_C} \\
B & \xrightarrow{f} & C \\
\end{array}
\]

This is precisely the defining diagram for a limit cone. Therefore, \((I, \xi)\) is a limit cone.

**Proof of 1209 (b).** Let \((L, \lambda)\) be a limit cone of the identity. The component \(\lambda_B\) of the natural transformation \(\lambda\) is a morphism from \(L\) to \(B\). In order for \(L\) to be an initial object, these morphisms must be unique.

Since \((L, \lambda)\) is a limit cone, for any cone \((A, \alpha)\) and any morphism \(f : B \to C\), there exists a unique morphism \(l_A : A \to L\) such that the following diagram commutes:

\[
\begin{array}{ccc}
A & \xrightarrow{l_A} & L \\
\downarrow{\lambda_B} & & \downarrow{\lambda_C} \\
B & \xrightarrow{f} & C \\
\end{array}
\]

In particular, there exists a unique map \(l_L : L \to L\) for the cone \((L, \lambda)\) such that the following diagram commutes:

\[
\begin{array}{ccc}
L & \xrightarrow{id_L} & L \\
\downarrow{\lambda_L} & & \downarrow{\lambda_L} \\
L & \xrightarrow{id_L} & L \\
\end{array}
\]

Since (467) commutes with \(id_L\) instead of \(l_L\), by uniqueness it follows that \(l_L = id_L\). Since, by the naturality of \(\lambda\), \(\lambda_L\) also satisfies this condition, \(\lambda_L = id_L\).

From the naturality of \(\lambda\), for any map \(f : L \to C\) it follows that the following diagram commutes:

\[
\begin{array}{ccc}
L & \xrightarrow{id_L} & C \\
\downarrow{\lambda_L} & & \downarrow{\lambda_C} \\
L & \xrightarrow{id_L} & C \\
\end{array}
\]

Therefore, \(f = \lambda_B\). Since \(f\) was an arbitrary morphism with domain \(L\), we conclude that \(L\) has a unique morphism to every object in \(C\). Hence, \(L\) is an initial object.
Example 1210. We will show that limits and colimits correspond to suprema and infima of partially ordered sets.

Let $(P, \leq)$ be a partially ordered set and $P$ be its corresponding thin skeletal category. The correspondence is discussed in theorem 1186 (a).

The image of a diagram $D : I \to P$ is a pair $(A, R)$, where $A$ is the set of objects $D(obj(I))$ in the image $D(I)$ and $R$ is a subrelation of $\leq$. This relation is reflexive, however it may not even be a preorder as shown in example 1147.

Conversely, every subset $A \subseteq P$ with a corresponding category $A$ is given by the diagram $I_A$, where the inclusion functor $I_A : A \to P$.

A cone of a diagram $D : I \to P$ is a lower bound of the set $D(obj(I))$. The relation induced by $D$ does not matter here.

Indeed, the cone $(x, R)$ in $P$ consists of morphisms with domain $x$. That is, $R$ is a subrelation of $\leq$ whose first component is $x$. Clearly then $x$ is a lower bound of the set $D(obj(I))$.

For the limiting cone $(y, S)$, it holds that $x \leq y$ for every other cone $(x, R)$. Both $R$ and $S$ are subrelations of $\leq$ with the same second components. Thus, $y$ is the greatest lower bound of $D(obj(I))$.

Dually, cocones are upper bounds and colimits are suprema.

As mentioned, the relation induced by the diagram $D$ does not actually matter. Therefore, we may choose, without loss of generality, $D$ to be a discrete category. It then follows that infima and suprema correspond to products and coproduct, however what we have shown here is more general.

Definition 1211. We will define limits and colimits of diagrams over infinite chains of integers. These notions predate limits and colimits, which explains why their names may seem inconsistent with other limits and colimits.

(a) Consider the category induced by the positive integers

\[
1 \xrightarrow{\ 2 \ } 2 \xrightarrow{\ 3 \ } 3 \xrightarrow{\ \cdots \ } n-1 \xrightarrow{\ n \ } n \xrightarrow{\ n+1 \ } n+1 \xrightarrow{\ \cdots \ }
\]

(469)

A colimit over a diagram of this shape is called a direct limit. In this context, the diagram itself is sometimes called a direct system.

(b) Now consider the opposite category

\[
\cdots \xrightarrow{\ n+1 \ } n \xrightarrow{\ n-1 \ } \cdots \xrightarrow{\ 3 \ } 2 \xrightarrow{\ 1 \ }
\]

(470)

Somewhat confusingly, a limit of a diagram of this shape is called an inverse limit. In this context, the diagram itself is sometimes called an inverse system.

Example 1212. We will list several examples of direct and inverse limits.

(a) Consider the chain of tuple vector spaces

\[
\mathbb{K} \xrightarrow{\ i_1^2 \ } \mathbb{K}^2 \xrightarrow{\ i_1^3 \ } \mathbb{K}^3 \xrightarrow{\ i_4 \ } \cdots \xrightarrow{\ i_{n-2} \ } \mathbb{K}^{n-1} \xrightarrow{\ i_{n-1} \ } \mathbb{K}^n \xrightarrow{\ i_{n+1} \ } \mathbb{K}^{n+1} \xrightarrow{\ \cdots \ }
\]

(471)
where \( \iota^m_n : \mathbb{K}^n \rightarrow \mathbb{K}^m \) is simply the canonical inclusion map.

Let \( \mathbb{K}_0^\infty \) be the vector space of all sequences in \( \mathbb{K} \) with only finitely many nonzero elements. For each positive integer \( n \), denote by \( \iota^n_0 : \mathbb{K}^n \rightarrow \mathbb{K}_0^\infty \) the canonical inclusion. Then \( (\mathbb{K}_0^\infty, \iota) \) is a cocone of (471). We will show that it is a colimit cocone, i.e. that \( \mathbb{K}_0^\infty \) is a direct limit of (471).

Let \( (A, \alpha) \) be another cocone. We want to define a linear map \( l_A : \mathbb{K}_0^\infty \rightarrow A \) so that the following diagram commutes:

\[
\begin{array}{ccc}
\mathbb{K}^n & \xrightarrow{i_n^m} & \mathbb{K}^m \\
\downarrow{\alpha_n} & & \downarrow{\alpha_m} \\
\mathbb{K}_0^\infty & \xrightarrow{l_A} & A \\
\end{array}
\]

(472)

For every vector \( x \in \mathbb{K}^m \), we must have

\( \alpha_n(x) = l_A(\iota^n_0(x)) \).

This implies the obvious definition where, given a vector \( x \in \mathbb{K}_0^\infty \) whose greatest nonzero element has index \( m \), we define

\( l_A(x) := \alpha_m(x) \).

This map is well-defined because the compatibility with inclusion maps guarantees that \( \alpha_{m+1}(x) = \iota_{m+1}(\iota^m_n(x)) \), i.e. the result obtained by using \( \alpha_m \) and \( \alpha_{m+1} \) is the same.

Therefore, \( (\mathbb{K}_0^\infty, \iota) \) is a direct limit of the diagram (471).

(b) Dually, consider the chain

\[
\cdots \xrightarrow{\pi^{n+1}_n} \mathbb{K}^n \xrightarrow{\pi^n_{n-1}} \mathbb{K}^{n-1} \xrightarrow{\pi^{n-2}_{n-1}} \cdots \xrightarrow{\pi^1_2} \mathbb{K}^3 \xrightarrow{\pi^2_1} \mathbb{K}^2 \xrightarrow{\pi^3_1} \mathbb{K} \quad (473)
\]

where

\[
\pi^m_n : \mathbb{K}^m \rightarrow \mathbb{K}^n \\
\pi^m_n(x_1, \ldots, x_n, x_{k+1}, \ldots, x_m) := (x_1, \ldots, x_n).
\]

Let \( \mathbb{K}^\infty \) be the vector space of all sequences in \( \mathbb{K} \) and, for each positive integer \( n \), define \( \pi^\infty_n \) as a truncation in the obvious way.

Then \( (\mathbb{K}^\infty, \pi^\infty) \) is a cone of (471). We will show that it is a limit cone, i.e. that \( \mathbb{K}^\infty \) is a direct limit of (471).
Let \((A, \alpha)\) be another cone. We want to define \(l_A : A \to \mathbb{K}^\infty\) so that the following diagram commutes:

\[
\begin{array}{ccc}
\mathbb{K}^n & \xrightarrow{\pi_n^m} & \mathbb{K}^m \\
\downarrow{\scriptstyle \pi_n} & & \downarrow{\scriptstyle \pi_m} \\
A & \xrightarrow{\sigma_n} & \mathbb{K}^\infty \\
\uparrow{\scriptstyle l_A} & & \uparrow{\scriptstyle \alpha_m} \\
\mathbb{K}^m & \xrightarrow{\pi_m} & \mathbb{K}^n
\end{array}
\]  

(474)

For every \(a \in A\), we must have

\[\alpha_n(a) = \pi_n^\infty(l_A(a)),\]

which implies the obvious definition where the \(n\)-th coordinate of the vector \(l_A(a)\) is

\[[l_A(a)]_n := [\alpha_n(a)]_n.\]

That is, \(l_A(a)\) is a sequence whose \(n\)-th coordinate is the \(n\)-th coordinate \(\alpha_n(a)\). This is well-defined because, for \(m > n\), we have

\[\pi_m^n \circ \alpha_m = \alpha_n.\]

(c) In concrete categories, inverse limits are generalizations of set intersections.

As an example, consider the chain of sets

\[\ldots \xrightarrow{\iota_{n+1}^n} A_n \xrightarrow{\iota_{n-1}^n} A_{n-1} \xrightarrow{\iota_{n-2}^n} \ldots \xrightarrow{\iota_3^1} A_3 \xrightarrow{\iota_2^1} A_2 \xrightarrow{\iota_1^2} A_1\]  

(475)

where \(A_{k+1} \subseteq A_k\) for every positive integer \(k\) and \(\iota_k^m : A_m \to A_k\) is simply the canonical inclusion map.

Then the intersection \(\bigcap_{k=1}^\infty A_k\) along with its inclusion maps is a limit of (475). This is a consequence of the discussion in example 1210.

We can replace the inclusion maps \(\iota_k^m\) with other injective functions, or even non-injective functions. In this case, we would obtain a “generalized intersection”.

**Definition 1213.** Fix an arbitrary category \(\mathcal{C}\). A **product** in \(\mathcal{C}\) is a limit over a diagram \(D : I \to \mathcal{C}\), whose domain \(I\) is a discrete category. Dually, a **coproduct** or **sum** in \(\mathcal{C}\) is a colimit of \(D\).

It will be convenient for us to speak about the product of an indexed family \(\{X_k\}_{k \in \mathcal{K}}\) of objects in \(\mathcal{C}\).

A cone with vertex \(A\) consists of morphisms with signatures

\[\alpha = \{\alpha_k : A \to X_k\}.\]

A cocone with vertex \(A\) consists of morphisms with signatures

\[\alpha = \{\alpha_k : X_k \to A\}.\]
A product cone \((\prod_{k \in \mathcal{K}} X_k, \pi)\) satisfies the following universal mapping property:

For every cone \((A, \alpha)\), there exists a unique morphism

\[ l_A : A \to \prod_{k \in \mathcal{K}} X_k, \]

such that the following diagram commutes:

\[ \begin{array}{ccc}
A & \xrightarrow{l_A} & \prod_{k \in \mathcal{K}} X_k \\
\downarrow{\pi_k} & & \downarrow{\prod_{k \in \mathcal{K}} \alpha} \\
X_k & \end{array} \]

We call the morphism \(\pi_k\) the **canonical projection** of the product onto \(X_k\), even though it may not be a surjective function.

From proposition 1208 it follows that the product (resp. coproduct) of an empty family is a terminal (resp. initial) object of \(\mathcal{C}\).

As in the case of general limits, we call \(\prod_{k \in \mathcal{K}} X_k\) the product of the family \(X\).

In the case of only two objects, their product is given by the following diagram:

\[ \begin{array}{ccc}
A & \xrightarrow{l_A} & X \times Y \\
\downarrow{\alpha_X} & \downarrow{\pi_X} & \downarrow{\pi_Y} \\
X & \end{array} \]

and their coproduct by

\[ \begin{array}{ccc}
A & \xrightarrow{\iota_A} & \coprod_{k \in \mathcal{K}} X_k \\
\downarrow{\alpha_X} & \downarrow{\iota_X} & \downarrow{\iota_Y} \\
X & \end{array} \]

**Proposition 1214.** The product in the category \(\mathcal{U}-\textbf{Set}\) of \(\mathcal{U}\)-small sets of a family \(\mathcal{A} = \{A_k\}_{k \in \mathcal{K}}\) is their Cartesian product \(\prod_{k \in \mathcal{K}} A_k\) and the coproduct is their disjoint union \(\coprod_{k \in \mathcal{K}} A_k\).

**Proof.**
\textbf{Proof for products.} Consider the Cartesian product

\[ L := \prod_{k \in \mathcal{K}} A_k = \left\{ f : A \to \bigcup_{k \in \mathcal{K}} A_k \mid \forall k \in \mathcal{K}. k \in A_k \right\}. \]

Since \( \mathcal{U} \) is a model of ZFC (with or without the axiom of infinity), hence the product is an object of \( \mathcal{U} \text{-} \textbf{Set} \). Define the projection morphisms

\[ \pi_k(f) := f(k). \]

Then \((L, \pi)\) is a cone for (some diagram forming) \( \mathcal{A} \). Let \((B, \beta)\) also be a cone and consider the following diagram:

\begin{equation}
\begin{array}{c}
B \\
\uparrow \beta_k \\
\prod_{k \in \mathcal{K}} A_k \\
\downarrow \pi_k \\
A_k
\end{array}
\end{equation}

(480)

We want to define a function \( l_B : B \to L \) such that

\[ \pi_k(l_B(b)) = \beta_k(b) \]

for every \( b \in B \) and \( k \in \mathcal{K} \). This is uniquely determined by \( b \) and \( k \), hence

\[ l_B(b) := \{ \beta_k(b) \}_{k \in \mathcal{K}}. \]

The cone \((B, \beta)\) was chosen randomly, and we showed that there exists a unique morphism \( l_B : B \to L \) such that (480) commutes. Therefore, \((L, \pi)\) is a categorical product of the family \( \mathcal{A} \).

\textbf{Proof for coproducts.} Now consider the disjoint union

\[ L := \biguplus_{k \in \mathcal{K}} A_k = \{(k, a) \mid k \in \mathcal{K} \text{ and } a \in A_k\} \]

and the inclusions

\[ l_k((k, a)) = a. \]

We must show that, for any other cocone \((B, \beta)\), the following diagram commutes:

\begin{equation}
\begin{array}{c}
B \\
\uparrow \beta_k \\
\biguplus_{k \in \mathcal{K}} A_k \\
\downarrow l_k \\
L
\end{array}
\end{equation}

(481)
Analogously to products, the condition

\[ l_B(t_k(a)) = \beta_k(a) \]

for every \( b \in B \) and \( k \in \mathcal{K} \) uniquely identifies the function

\[ l_B(k,a) := \beta_k(a). \]

\[ \square \]

[Lei16] **Definition 1215.** Consider the index category

\[ \bullet \quad \rightarrow\longrightarrow\rightarrow \bullet \]  

(482)

A diagram indexed by (482) is sometimes called a **fork**. A limit of a fork is called an **equalizer** and a colimit — a **coequalizer**.

Note that in definition 1148 we defined commutativity only when at least one path is nontrivial. That is, (482) cannot commute by definition since all of its paths have length either 0 or 1.

We will describe equalizers in more detail. Fix a fork in \( \mathbf{C} \) which we will denote by

\[ X \overset{s}{\underset{t}{\rightrightarrows}} Y \]  

(483)
A cone with vertex $A$ over this diagram has a single morphism $f : A \to X$ such that the following diagram commutes:

$$A \xrightarrow{f} X \xrightarrow{s} Y$$  \hspace{1cm} (484)

This is equivalent to requiring

$$s \circ f = t \circ f.$$

An equalizer limit cone $(L, \iota)$ satisfies the following universal mapping property:

For every cone $(A, f)$, $f$ uniquely factors through $L$. That is, there exists a unique morphism $l_A : A \to L$ such that the following diagram commutes:

$$A \xleftarrow{l_A} X \xrightarrow{s} Y$$  \hspace{1cm} (485)

The requirement that (485) commutes does not mean that $p = q$. As discussed in definition 1148, for commutative diagrams, we only consider a pair of paths if at least one of them is nontrivial. We made this requirement in order to allow parallel morphisms.

Note how we interchanged the notation for the projections and inclusions compared to definition 1213 — $(L, \iota)$ is a limit of a fork, while $(L, \pi)$ is a colimit. The reason for this is that equalizers are usually canonical inclusions, while coequalizers are projections.

**Proposition 1216.** If $(L, \iota)$ is an equalizer in the category $C$, then $\iota$ is a monomorphism. Dually, if $(L, \pi)$ is a coequalizer, then $\pi$ is an epimorphism.

**Proof.** Fix an equalizer cone $(L, \iota)$ of the fork (483). Fix any object $A$ in $C$ and any two parallel morphisms $a_1, a_2 : A \to X$ such that

$$\iota \circ a_1 = \iota \circ a_2.$$

Then

$$g \circ \iota \circ a_1 = h \circ \iota \circ a_2,$$

and thus $\iota \circ a_1 = \iota \circ a_2$ is the morphism of a cone. Hence, both $(A, \iota \circ a_1)$ and $(A, \iota \circ a_2)$ are
cones and thus $a_1$ and $a_2$ are the unique maps such that the following diagram commutes:

\[
\begin{array}{c}
A \\
\downarrow a_1 \\
L & \xrightarrow{\iota} & B & \xrightarrow{g} & C \\
\downarrow a_2 \\
A
\end{array}
\]

(488)

Therefore, $a_1 = a_2$ and, since $a_1$ and $a_2$ were arbitrary, it follows that $\iota$ is a monomorphism.

Dually, if $(L, \pi)$ is a coequalizer colimit cocone in $\mathbf{C}$, then by proposition 1203, $(L, \pi^{\text{op}})$ is an equalizer in $\mathbf{C}^{\text{op}}$. Hence, $\pi^{\text{op}}$ is a monomorphism, and by proposition 1127, $\pi$ is an epimorphism. □

**Example 1217.**

(a) For a pair of functions $s, t : X \to Y$ in $\mathbf{Set}$, define the set

\[ E = \{x \in X \mid s(x) = t(x)\}. \]

With the inclusion map $\iota : E \to X$, this is an equalizer cone for $s$ and $t$.

We will now prove that it is a limit cone. The pair $(E, \iota)$ is obviously a cone. For any other cone $(A, f)$, we must find a map $l_A : A \to E$ such that the following equalizer diagram commutes:

\[
\begin{array}{c}
A \\
\downarrow f \\
X & \xrightarrow{s} & Y \\
\downarrow \iota \\
E
\end{array}
\]

(489)

In order for $(A, f)$ to be a cone, the image $f[A]$ must be a subset of $E$. In order for (489) to commute, it only makes sense to take $l_A$ to be $f$ with its codomain restricted to $E$. It follows that $(E, \iota)$ is a limit cone.

(b) The coequalizer of $s, t : X \to Y$ is more nuanced. Suppose that $(L, \pi)$ is a colimit cocone and $(A, f)$ is any cocone for the coequalizer diagram

\[
\begin{array}{c}
X & \xrightarrow{s} & Y \\
\downarrow \pi \\
L
\end{array}
\]

(490)

In order for $(A, f)$ to be a cocone, for every $x \in X$, we must have $f(s(x)) = f(t(x))$. Thus, $A$ must be a partition of $Y$ in a way such that $s(x)$ and $t(x)$ belong to the same coset if and only if $f(s(x)) = f(t(x))$. Outside the images of $s$ and $t$, $f$ is free to take any value.
Let $\sim$ be the smallest equivalence relation on $Y$ such that $s(x) \sim t(x)$ for every $x \in X$. Explicitly, this is the **equivalence closure** of the relation
\[
\{(s(x), t(x)) \mid x \in X\}.
\]

Consider the partition $Y/\sim$ with the projection map
\[
\begin{align*}
\pi &: Y \to Y/\sim \\
\pi(y) &: = [y]
\end{align*}
\]

This is a cocone. Furthermore, it is a colimit cocone because, for any other cocone $(A, f)$, we can define $l_A([y]) := f(y)$ so that (490) commutes.

Coequalizers mostly make sense in the context of **quotient groups**, where partitions are especially well-behaved and admit a much simpler description. See **definition 455 (i)**.
15. Order theory

Orders are special binary relations which, surprisingly, are used to compare elements in a set. Order theory studies pairs \( X = (X, \leq) \), where \( X \) is a set and \( \leq \) is a preorder.

We denote orders using symbols rather than letters because it is customary to write orders using infix notation, e.g. \( a \leq b \) rather than \((a, b) \in \leq\).

![Figure 31: Hierarchy of ordered sets]

Section 13.5 (Well-ordered sets) is focused on set theory, and hence we have included it in section 13 (Set theory) rather than here.

General (semi)lattices also admit algebraic definitions, however these algebraic descriptions have some drawbacks:

- There is no general way to extend algebraic operations from finitary to infinitary. This can sometimes be circumvented if the order happens to carry more information about the partially ordered set than the algebraic operations — see proposition 1255 (e).

- We often implicitly rely on the order structure, for example in the definition for Heyting algebras.
15.1. Preordered sets

**Definition 1218.** A **preordered set** is a set \( P \) endowed with a reflexive and transitive binary relation \( \leq \). The relation itself is called a **preorder**.

It is conventional to use the same symbol \( \leq \) as for partial orders, however the lack of antisymmetry may be confusing — see example 1225.

We define \( \geq \) as the inverse relation of \( \leq \).

Preordered sets have the following metamathematical properties:

(a) Consider a first-order language \( \mathcal{L} \) with two infix binary predicate symbols — \( \leq \) and \( \geq \). The theory of preordered sets is a first-order theory in \( \mathcal{L} \) consisting of the axioms (329) and (333) for \( \leq \) and the compatibility axiom

\[
(\xi \leq \eta) \leftrightarrow (\eta \geq \xi).
\]  

(b) A homomorphism from \( (P, \leq_P) \) to \( (Q, \leq_Q) \) is, explicitly, a function \( f : P \to Q \) such that

\[
x \leq_P y \text{ implies } f(x) \leq_Q f(y).
\]  

These are precisely the nonstrict monotone maps.

(c) Since the theory contains only positive formulas over a language with no functional symbols, any subset \( A \) of the domain of a preordered set \( P \) becomes a preordered set with the induced preorder \( \leq_A \) defined as the **restriction** of \( \leq_P \) to only elements of \( A \).

(d) The trivial preordered set is the empty set. See remark 827 regarding allowing empty sets as first-order structures.

(e) We denote the category of \( \mathcal{U} \)-small models for the theory of preordered sets by \( \mathcal{U} \text{-PreOrd} \). This category is equivalent to that of \( \mathcal{U} \)-small thin categories — see theorem 1186 (a).

(f) We define the **opposite preordered set** of \( (P, \leq) \) as \( (P, \geq) \).

The **principle of duality** states that the formula \( \varphi \) is derivable in the theory of preordered sets if and only if the opposite formula \( \varphi^{op} \), in which we swap all instances of \( \leq \) and \( \geq \), is also derivable. Observe that \( \varphi \) is satisfied in a preordered set if and only if \( \varphi^{op} \) is satisfied in the opposite preordered set.

There is a actually a very simple proof. If \( \varphi \) is derivable in the theory, it is satisfied by every preordered set. Let \( P \) be a preordered set. Then \( \varphi \) is satisfied in both \( (P, \leq) \) and its opposite \( (P, \geq) \). But if \( \varphi \) is valid for the opposite preordered set \( (P, \geq) \), its opposite formula \( \varphi^{op} \) is valid for the the double opposite preordered set, which is \( (P, \leq) \). Since \( (P, \leq) \) was chosen arbitrarily, the opposite formula \( \varphi^{op} \) is valid for every preordered set and, so it belongs to the theory of preordered sets.
The opposite of the opposite formula of \( \varphi \) is obviously \( \varphi \). The actual replacement can be formalized by performing the \textit{simultaneous substitution}

\[
\varphi^{\text{op}} := \varphi[\xi_1 \leq \eta_1 \mapsto \xi_1 \geq \eta_1, \quad \xi_1 \geq \eta_1 \mapsto \xi_1 \leq \eta_1, \\
\vdots \quad \vdots \\
\xi_n \leq \eta_n \mapsto \xi_n \geq \eta_n, \quad \xi_n \geq \eta_n \mapsto \xi_n \leq \eta_n]
\]

for all pairs \((\xi_k, \eta_k)\) of free variables in \( \varphi \).

Another form of this duality is formalized in \textit{Theorem 1186} (Ordered sets as categories).

\textbf{Definition 1219.} A preordered set \( P \) is called an \textbf{upward/right directed set} if every finite subset of \( P \) has an \textbf{upper bound}, i.e. for all \( x, y \in P \) there must exist \( z \in P \) such that \( x \leq z \) and \( y \leq z \). We do not care how many upper bounds exist and how they are related, we simply need one upper bound to exist for every pair of elements of \( P \).

Dually, \( P \) is a \textbf{downward/left directed set} if every two elements have a lower bound.

Directed sets are used to define nets in topological spaces, see \textit{Definition 272}.

\textbf{Definition 1220.} A subset \( A \) of a preordered set \((P, \leq)\) is called \textbf{cofinal} if for every \( x \in P \) there exists some \( y \in A \) such that \( x \leq y \).

\textbf{Example 1221.} We list several examples of \textbf{cofinal} and non-cofinal sets.

- In a finite set like \( \{0, 1, 2\} \), the set \( \{2\} \) containing the maximum is cofinal. This is generalized by \textit{Proposition 1233}.

- Consider the set \( \mathbb{Z} \) of integers. Clearly the set \( 2\mathbb{Z} \) of even integers is cofinal. This is generalized by \textit{Proposition 1247}.

- Cofinal sets are important in topology because it is used to define \textbf{convergence of nets}.

- \textbf{Regular cardinals} are equal to their own \textbf{cofinality}.
15.2. Partially ordered sets

Preordered sets are simple to define and arise naturally (e.g. definition 918), but they require uniqueness considerations, as discussed in example 1225, and we want to avoid them. Proposition 1226 shows that the result obtained from this factorization is a partially ordered set, which this section is dedicated to.

Definition 1222. A partial order on a set \( P \) can be defined in the following equivalent ways:

(a) A preorder \( \leq \) on \( P \) such that \( \leq \) is antisymmetric in addition to being reflexive and transitive. This definition is the more common one. If we wish to distinguish it from the other definition, we call such a relation a nonstrict partial order.

(b) An irreflexive and transitive binary relation \( < \) on \( P \). This relation is called a strict partial order.

If both relations are present, in order for them to be equivalent, \( \leq \) must be the union of \( < \) and the diagonal \( \Delta \). This condition corresponds to the following axiom:

\[
(x \leq y) \leftrightarrow (x < y) \lor (x = y).
\] (493)

By adding \( \cdot \land \lnot(x = y) \) to both sides of (493), using theorem 1279 (De Morgan’s laws) and taking irreflexivity of \( < \) into account, we obtain

\[
(x < y) \leftrightarrow ((x \leq y) \land \lnot(x = y)).
\] (494)

A set \( P \) with any of the two types of partial orders is called a partially ordered set or simply a partially ordered set.

The elements \( x, y \in P \) are called comparable if either \( x \leq y \) or \( y \leq x \). That is, they are comparable if they are related by \( \leq \).

Partially ordered sets have the following metamathematical properties:

(c) Since we can interdefine nonstrict and strict orders, it makes little sense to study different theories for the two.

In order to define the theory of partially ordered sets, we extend the language of the theory of preordered sets with two infix binary predicate symbols — \( < \) and \( > \). We then add to the axioms of the theory (954 (i)) for \( \leq \) and either of the compatibility conditions (493) or (494) (it is unnecessary to add both).

We can also add (954 (g)) and (954 (j)) for \( < \), but that would be redundant.

(d) We have two types of first-order homomorphisms between partially ordered sets. The homomorphisms are often called order homomorphisms or monotone maps.

A nonstrict monotone map from \( (P, \leq_P) \) to \( (Q, \leq_Q) \) is a homomorphism for the theory containing only \( \leq \) and \( \geq \):

\[
x <_P y \text{ implies } f(x) \leq_Q f(y).
\] (495)
Nonstrict homomorphisms are used extensively in the theory of partially ordered sets, in particular in lattice theory.

A strict monotone map is instead a homomorphism for the full theory:

\[ x <_P y \implies f(x) <_Q f(y). \]  \hfill (496)

Strict homomorphisms are used in the theory of totally ordered sets, in particular for well-ordered sets and ordinals.

See Proposition 1228 for the converse implications to those in (495) and (496).

In particular, if \( P \) is the preordered set of positive integers, then we speak of monotone sequences

\[ \{x_k\}_{k=1}^\infty, \]  \hfill (497)

where \( x_{k-1} \leq_Q x_k \) for all \( k \in 1, 2, 3, \ldots \).

To elaborate, order homomorphisms can either be strict or nonstrict depending on whether we include \(<\) and \(>\) as predicate symbols in the language of the theory. We will use the convention that even when the language does include \(<\) and \(>\), by “order homomorphism” we will understand nonstrict monotone.

Furthermore, Proposition 1243 shows that for totally ordered sets, strict order homomorphisms are precisely the strong order homomorphisms (in the sense of Remark 857).

(e) As for preordered sets, any subset of a partially ordered set is itself a partially ordered set.

(f) The trivial partially ordered set is the empty set (see Remark 827 regarding allowing empty sets as first-order structures).

(g) We denote the category of \( \U \)-small models by \( \U\text{-Pos} \). It is a full subcategory of the category \( \U\text{-PreOrd} \) of \( \U \)-small preordered sets.

The category \( \U\text{-Pos} \) is equivalent to that of \( \U \)-small thin skeletal categories — see Theorem 1186(b).

(h) The principle of duality for preordered sets holds for partially ordered sets if we also swap \(<\) and \(>\).

Proof.

Proof that 1222 (a) implies 1222 (b). Let \( \leq \) be a nonstrict partial order. We will show that \(<\) is a strict partial order.

Proof of transitivity. The relation \(<\) is transitive. To see this, let \( x < y \) and \( y < z \). In particular, \( x \leq y \) and \( y \leq z \). From transitivity, we have \( x \leq z \).

Additionally, \( x \neq y \) and \( y \neq z \). Assume that \( x = z \). From reflexivity of \( \leq \) we have \( z \leq x \) and, since \( y \leq z \), from transitivity we obtain \( y \leq x \). But since \( x \leq y \), from the antisymmetry of \( \leq \), we have \( x = y \), which contradicts the assumption that \( x < y \).

Therefore, \( x < z \).
Figure 32: A drawing of the Hasse diagram (499)

**Proof of irreflexivity.** Follows directly from reflexivity of ≤ and the compatibility condition. Since the right side is false, the left side \( x < x \) is also false.

**Proof that 1222 (b) implies 1222 (a).** Let \( < \) be a strict partial order. We will show that \( \leq \) is a nonstrict partial order.

**Proof of reflexivity.** Fix \( x \in P \) and assume that \( x \not\leq x \). Then \( x \neq x \) which contradicts the reflexivity of equality. Hence, \( x \leq x \).

**Proof of antisymmetry.** Let \( x \leq y \) and \( y \leq x \), that is, either \( x = y \) or both \( x < y \) and \( y < x \) hold. Assume the latter. By the transitivity of \( \leq \), we have \( x < x \), which contradicts the irreflexivity of \( < \). Hence, \( x = y \).

**Proof of transitivity.** Let \( x \leq y \) and \( y \leq z \). Then we have four cases depending on which of \( x \), \( y \) and \( z \) are equal. Since both relations \( < \) and \( = \) are transitive, it follows that in all four cases \( x \leq z \).

**Definition 1223.** It is usually easier to define small finite partially ordered sets by drawing graphs than by enumerating all relation pairs. Let \((P, \leq)\) be a finite partially ordered set. The relation \( \leq \) may also be regarded as the set of edges of a directed graph. The graph \((P, \text{red}^{T}(\leq))\), whose edges are the transitive reduction of \( \leq \), is called the Hasse graph or Hasse diagram of \( P \).

The term “Hasse diagram” is usually associated with drawings. By convention, no arrowheads for denoting directions are drawn on the Hasse graph despite the graph being directed; instead, edges always point upwards. See example 1224.

**Example 1224.** Consider the partial order over \( \{a, b, c, d, e\} \) defined via

\[
\begin{align*}
a &\leq c, & a &\leq d, & a &\leq e, & b &\leq d, & b &\leq e, & d &\leq e.
\end{align*}
\]  

(498)

The corresponding Hasse graph includes only the underlined edges. The rest of the edges can be restored from transitivity. In this case, the Hasse graph has edges

\[
\{a \to c, a \to d, b \to d, d \to e\}
\]  

(499)

**Example 1225.** Consider the preordered set \( P \) in fig. 33 in which \( b \leq c \) and \( c \leq b \), but \( b \neq c \). We cannot properly draw a Hasse diagram because we have the restriction that \( c \) is drawn (strictly) higher than \( b \) if \( c > b \) and that \( c \) is drawn lower than \( b \) if \( c < b \). We face a
similar problem formally, for example in the definition of a Lindenbaum-Tarski algebra of a logical theory, where the preorder $\vdash$ allows $\varphi \vdash \psi$ and $\psi \vdash \varphi$, but still $\varphi \neq \psi$. Thus, we have nonuniqueness — every tautology is a largest element with respect to $\vdash$, while we want to have a single largest element for the sake of building a tidier theory.

If we are only interested in members of $P$ up to the equivalence (500), it is easy to factor $P$ by the equivalence relation (500) and obtain a partially ordered set. In the language of graph theory, if we have directed cycles that we may wish to avoid, we can contract each directed cycle into a single vertex, at which point the graph becomes acyclic. This corresponds to graph condensation.

The formulation and proof of correctness of this process can be found in proposition 1226 and an example can be found in fig. 33.

**Proposition 1226.** Let $(P, \leq)$ be a preordered set. Define the relation $\cong$ by

\[ x \cong y \text{ if and only if } x \leq y \text{ and } y \leq x. \quad (500) \]

That is, $\cong$ is the intersection of the relation $\leq$ with its inverse.

Since $\cong$ is an equivalence relation we can for the the quotient set $P/ \cong$. Define the relation $\leq$ on this quotient set by

\[ [x] \leq [y] \text{ if and only if } x \leq y. \]

The pair $(P/ \cong, \leq)$ is then a partially ordered set.

**Proof.** The relation $\leq$ is well-defined. Indeed, let $x \cong x'$ and $y \cong y'$, that is, both $x \leq x'$ and $x' \leq x$ and similarly for $y$. If $x \leq y$, from transitivity $x \leq y \leq y'$. But $x' \leq x$, hence $x' \leq y'$.

It is then clear that $\leq$ is a partial order because it inherits reflexivity and transitivity from $\leq$ and antisymmetry is imposed by taking quotient sets — equality in $P/ \cong$ holds precisely when $\cong$ holds in $P$.

Thus, $(P/ \cong, \leq)$ is indeed a partially ordered set. \qed

**Proposition 1227.** For any two preordered sets $(P, \leq_P)$ and $(Q, \leq_Q)$, every order embedding of $P$ into $Q$ is a strict order embedding.

Compare this to proposition 1242

**Proof.** Let $f : P \to Q$ be an order embedding.
Let \( x <_P y \) for some members \( x \) and \( y \) of \( P \). Since \( f \) is an order homomorphism, we have \( f(x) \leq_Q f(y) \). Since it is also injective, \( f(x) = f(y) \) implies \( x = y \), which contradicts our previous assumption.

Therefore, \( f(x) <_Q f(y) \) and \( f \) is a strict order homomorphism. \( \square \)

**Proposition 1228.** For any (even nonstrict) monotone map \( f : P \to Q \) between partially ordered sets, if \( x \) and \( y \) are comparable elements of \( P \) we have

\[
x <_P y \quad \text{if and only if} \quad f(x) <_Q f(y).
\]

(501)

**Proof.** Let \( f(x) <_Q f(y) \) and suppose that \( x \geq y \). Since \( f \) is a monotone map, we have \( f(x) \geq_Q f(y) \), which is a contradiction. \( \square \)

**Definition 1229.** We introduce the following terminology for extremal elements of a partially ordered set \( P \). Analogous definition can be given for preordered sets, but the nonuniqueness problems outlined in example 1225 highlight that there are sometimes difficulties in doing, so.

The notions on the left and on the right are dual, but we discuss both nonetheless.

(a) An **upper bound** for the set \( A \subseteq P \) is an element \( x_0 \in P \) such that \( x \leq x_0 \) for every \( x \in A \). Note that \( x_0 \) does not in general belong to \( A \). An upper bound is called **strict** if it does not belong to \( A \).

If \( A \) has at least one upper bound, it is called **bounded from above**.

Every element is vacuously an upper bound of \( A = \emptyset \).

In fig. 32, the set \( A = \{a, b\} \) is bounded from above by both \( d \) and \( e \), but the entire partially ordered set has no upper bound.

Dually, \( x_0 \in P \) is a **lower bound** of \( A \) if \( x_0 \leq x \) for every \( x \in A \). A lower bound is called **strict** if it does not belong to \( A \).

If \( A \) has a lower bound, it is called **bounded from below**.

If \( A \) is bounded both from below and from above, we say that \( A \) is **bounded**.

Every element is vacuously a lower bound of \( A = \emptyset \). Hence, the empty set is bounded.

In fig. 32, the entire partially ordered set has no lower bound. The set \( A = \{c, d\} \) is bounded from below by \( a \), but not from above, hence \( A \) is not bounded.

The member \( x_0 \in A \) is a **minimal element** of \( A \) if for every \( x \in A \) such that \( x \leq x_0 \) we have \( x = x_0 \).

The empty set cannot have maximal or minimal elements because it has no members.

In fig. 32, the entire partially ordered set has two incomparable maximal elements — \( c \) and \( e \).

(b) A **maximal element** for the set \( A \subseteq P \) is a member \( x_0 \) of \( A \) such that there is no greater element in \( A \) than \( x_0 \). More precisely, \( x_0 \) is a maximal element of \( A \) if for every element \( x \in A \) such that \( x \leq x_0 \) we have \( x = x_0 \).

In fig. 32, the entire partially ordered set has two incomparable maximal elements — \( a \) and \( b \).
(c) The **maximum** or greatest element of $A \subseteq P$, if it exists, is an upper bound of $A$ that belongs to $A$. A maximum is necessarily a maximal element because $x_0 \leq x$ only holds for $x = x_0$, which also demonstrates unicity of $x_0$. See example 1230 for a unique maximal element that is not a maximum.

The **minimum** of $A$, also called smallest element or least element, is a lower bound that belongs to $A$. The empty set cannot have a maximum or minimum because it has no members. In fig. 32, the entire partially ordered set has no maximum, but the set $A = P \setminus \{b\} = \{a, b, d, e\}$ has $a$ as its minimum.

The **supremum** $\sup A$ of $A \subseteq P$, if it exists, is its least upper bound of $A$, i.e. the minimum of the set of its upper bounds. In fig. 32, the entire partially ordered set has no upper bound, so it cannot possibly have a supremum. Obviously every supremum is a maximum, but the converse is not true. We already noted that both $d$ and $e$ are upper bounds of the set $\{a, b\}$ and since $d \leq e$, we conclude that $d$ is the supremum of $\{a, b\}$, yet the set $\{a, b\}$ has no maximum.

Dually, the minimum of $P$ is usually denoted by $\perp$ and called the global minimum or bottom element of $P$.

The supremum of the empty set is the least of the upper bounds of the empty set, i.e. the minimum of $P$, which is $\perp$. In conclusion,

$$\perp = \min P = \inf P = \sup \emptyset.$$  

If $\perp$ exists, we say that the partially ordered set $P$ itself is bounded from below.

**(d)** The supremum $\sup A$ of $A \subseteq P$, if it exists, is its least upper bound of $A$, i.e. the minimum of the set of its upper bounds. In fig. 32, the entire partially ordered set has no upper bound, so it cannot possibly have a supremum. Obviously every supremum is a maximum, but the converse is not true. We already noted that both $d$ and $e$ are upper bounds of the set $\{a, b\}$ and since $d \leq e$, we conclude that $d$ is the supremum of $\{a, b\}$, yet the set $\{a, b\}$ has no maximum.

If it exists, the maximum of the entire partially ordered set $P$ is usually denoted by $\top$ and called the global maximum or top element of $P$. Since $\top$ is the maximum of $P$, it is also the supremum of $P$. Since every member of $P$ is a lower bound of $\emptyset$, the greatest lower bound is the maximum of $P$.

In conclusion,

$$\top = \max P = \sup P = \inf \emptyset.$$  

If $\top$ exists, we say that the partially ordered set $P$ itself is bounded from above.

**Definition 1231.** A **chain** in a partially ordered set is a subset in which every two elements are comparable. The length of a chain $A$ is the cardinal number $\text{card}(A) - 1$. The length of
A partially ordered set, if it exists, is the maximum among the lengths of all its chains. The length of a partially ordered set is also called its **height** because of how Hasse diagrams are drawn.

An **antichain** is a subset in which no two elements are comparable. The **width** of a partially ordered set, if it exists, is the maximum among the cardinalities of its antichains.

The height of the partially ordered set in fig. 32 is 2 and it is reached by the chains \{a, d, e\} and \{b, d, e\}. The width is 2 and is reached by \{a, b\}, \{b, e\}, \{c, d\} and \{c, e\}.

**Definition 1232.** Fix a partially ordered set \((P, \leq)\). For any \(a, b \in P\) with \(a \leq b\), we define the following related partially ordered sets:

(a) The **open rays**, also called the **open initial segment** and **open final segment**, are defined as

\[
(a, \infty) := \{x \in P \mid b > a\} =: P_{>a}, \\
(-\infty, b) := \{x \in P \mid x < b\} =: P_{<b}.
\]

The notation on the left assumes that the sentinel symbols \(\infty\) and \(-\infty\) are adjoined to \(P\) as in the case of the extended real numbers. This convention is widespread for unbounded ordered rings representing numbers — for example \(\mathbb{Z}, \mathbb{Q}\) and \(\mathbb{R}\). The term “ray” is used in this context due to the connection with geometric rays.

The notation on the right is more general and is widespread for abstract partial orders, most notably **well-ordered sets**. In the latter context, they are usually referred to as “initial/final segments”.

The **closed rays** and **closed initial/final segments** are defined analogously as

\[
[a, \infty) := \{x \in P \mid x \geq a\} =: P_{\geq a}, \\
(-\infty, b] := \{x \in P \mid x \leq b\} =: P_{\leq b}.
\]

We implicitly assume that \(a \leq b\), but this is not strictly necessary — \([a, b]\) is an empty set otherwise.

(b) The **closed interval** with endpoints \(a\) and \(b\) is

\[
[a, b] := \{x \in P \mid a \leq x \leq b\} = P_{\geq a} \cap P_{\leq b}.
\]

We implicitly assume that \(a \leq b\), but this is not strictly necessary — \([a, b]\) is an empty set otherwise.

(c) The **open interval** with endpoints \(a\) and \(b\) is

\[
(a, b) := \{x \in P \mid a < x < b\} = P_{>a} \cap P_{<b}.
\]

We implicitly assume that \(a < b\), but this is also not strictly necessary.

(d) The **half-open intervals** are

\[
(a, b] := \{x \in P \mid a < x \leq b\} = P_{>a} \cap P_{\leq b}, \\
[a, b) := \{x \in P \mid a \leq x < b\} = P_{\geq a} \cap P_{<b}.
\]
**Proposition 1233.** Let \((P, \leq)\) be a *bounded from above* totally ordered set and let \(A \subseteq P\). Then \(A\) is *cofinal* if and only if it contains the *top element* \(\top\).

*Compare this result with Proposition 1247.*

**Proof.**

**Proof of sufficiency.** Let \(A\) be a cofinal set. Then \(A\) must contain an element \(x\) such that \(\top \leq x\). But \(\top\) is a maximum and hence \(x = \top\) and thus \(\top \in A\).

**Proof of necessity.** Let \(A\) be a set containing \(\top\). Then for any \(x \in P\) we have \(x \leq \top\) and hence \(A\) is cofinal. \(\square\)

**Definition 1234.** Let \((P, \leq_P)\) and \((Q, \leq_Q)\) be partially ordered sets.

The *lexicographic order* on \(P \times Q\), also known as the *dictionary order*, is defined as

\[(a, b) < (c, d) \text{ if and only if } \left( a <_P c \text{ or } \left( a = c \text{ and } b <_Q d \right) \right). \tag{502}\]

The *reverse lexicographic order* is

\[(a, b) < (c, d) \text{ if and only if } \left( b <_Q d \text{ or } \left( b = d \text{ and } a <_P c \right) \right). \tag{503}\]

The lexicographic order on \(P \times Q\) inherits some important properties from \((P, P)\) and \((Q, \leq_Q)\) as can be seen in propositions 1001, 1236 and 1248.

We can use natural number recursion to extend this to arbitrary \(n\)-tuples — see example 1235 for an example.

**Example 1235.** A key example for *lexicographic ordering* is real-world dictionary like a thesaurus. It is obvious that “homomorphism” should come after “axiom” and the lexicographic order on a Cartesian power of Latin alphabets suggests that “homeomorphism” should also come after “homeomorphic”.

A slightly more relevant example for mathematics is the lexicographic ordering

\[AB < AD < BC < CD\]

of the names of the edges of a rectangle. The reverse lexicographic ordering is

\[AB < BC < AD < CD.\]

For the sides of a triangle, we have \(AB < AC < BC\) for both orderings.

The edges of the graph in eq. (552) are numbered in lexicographic order, which also happens to be the reverse lexicographic order.

**Proposition 1073** and **proposition 1075** contain more interesting applications of lexicographic orders.

**Proposition 1236.** The *lexicographic* and *reverse lexicographic* orders are *strict partial order* relations.

*Compare this result to proposition 1248 and proposition 1001.*
Proof.

**Proof of irreflexivity.** Trivial.

**Proof of transitivity.** Let \(\prec\) be a lexicographic order on \(P \times Q\). If \((a, b) \prec (c, d)\) and \((c, d) \prec (e, f)\), then:

- If \(a < c\), then \(a < c \leq e\) and thus \((a, b) \prec (e, f)\).
- If \(a = c\) and \(b < d\), then \(a \leq e\) and \(b < d \leq f\) and thus \((a, b) \prec (e, f)\).

The proof for the reverse lexicographic order is analogous.

**Definition 1237.** An endofunction on a partially ordered set \((P, \leq)\) is said to be inflationary if \(x \leq \text{cl}(x)\) for every \(x \in P\).

**Definition 1238.** Let \((P, \leq)\) be a partially ordered set. We say that the function \(\text{cl} : P \to P\) is a closure operator if it is inflationary, idempotent and nonstrictly order-preserving.

We say that \(x\) is closed with respect to \(\text{cl}\) if \(x = \text{cl}(x)\).

**Proposition 1239.** For any closure operator \(\text{cl} : P \to P\) on any partially ordered set \((P, \leq)\) and any \(x \in P\) it holds that \(\text{cl}(x)\) is the smallest closed element of \(P\) containing \(x\).

**Proof.** Since \(f\) is inflationary, it is clear that \(x \leq \text{cl}(x)\).

Let \(y \in P\) be a closed element such that \(x \leq y \leq \text{cl}(x)\). From the monotonicity of \(f\) we have that

\[
\text{cl}(x) \leq \underbrace{\text{cl}(y)}_{\text{cl}(x)} \leq \text{cl}(\text{cl}(x))
\]

Therefore, \(y = \text{cl}(x)\).

**Theorem 1240** (Zorn’s lemma). If every chain in a partially ordered set has an upper bound, then the entire set has a maximal element.

Zorn’s lemma is usually stated and used only in a lattice of sets, however it is a more general statement in order theory.

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (k).

**Proof.**

**Proof that the axiom of choice implies Zorn’s lemma.** Let \((P, \leq)\) be a partially ordered set in which every chain has an upper bound. Aiming at a contradiction, suppose that \(P\) has no maximal element.

Denote by \(\mathcal{C}\) the set of all chains of \(P\) and define the multi-valued map \(F : \mathcal{C} \to P\) that assigns to each chain the set of all strict upper bounds.

Since \(P\) has no maximal element, every chain has a strict upper bound. That is, the function \(F\) is a total multi-valued map. By theorem 986 (Multi-valued selection existence), there exists a single-valued selection \(f : \mathcal{C} \to P\) of \(F\). We have indirectly used the axiom of choice via theorem 986 (Multi-valued selection existence).
By theorem 1029 (Hartogs’ lemma), there exists a smallest ordinal $\alpha$ such that no function from $\alpha$ to $P$ is injective. Using theorem 1010 (Bounded transfinite recursion), we can define

$$g : \alpha \to A$$

$$g(\beta) := f(\{g(\gamma) | \gamma < \beta\}).$$

By construction, for every $\beta < \alpha$ the value $g(\beta)$ is a strict upper bound of the set $\{g(\gamma) | \gamma < \beta\}$. Hence, the function $f$ is injective, which directly contradicts our choice of $\alpha$.

The obtained contradiction shows that $P$ has a maximal element.

**Proof that Zorn’s lemma implies axiom of choice.** Let $\mathcal{A}$ be a family of nonempty sets. Let $\mathcal{F}$ be the set of all partial single-valued functions from $\mathcal{A}$ to $\bigcup \mathcal{A}$ with the subset ordering. That is, $f \leq g$ if $\text{dom}(f) \subseteq \text{dom}(g)$ for $f, g \in \mathcal{F}$.

Clearly every chain has a maximum - a total single-valued function. Then $\mathcal{F}$ itself has a maximal element by Zorn’s lemma. This maximal element is necessarily a total function because otherwise it would not be maximal.

Then this is the desired choice function for the family $\mathcal{A}$. \qed
15.3. Totally ordered sets

Definition 1241. We say that a partially ordered set is **totally ordered** if either the nonstrict order \( \leq \) is **total** or if the strict order \( < \) is **trichotomic**.

The theory, homomorphisms and category \( \text{Tos} \) are obtained analogously to definition 1222, but with either of these additional axiom sets.

For a fixed Grothendieck universe, the category of \( \mathcal{U} \)-\( \text{Tos} \) is isomorphic to that of \( \mathcal{U} \)-small thin skeletal connected categories — see theorem 1186 (c).

Proof. Equivalence between nonstrict and strict total orders follows directly from the compatibility condition (493).

Proposition 1242. Let \( \left( P, \leq_P \right) \) and \( \left( Q, \leq_Q \right) \) be totally ordered sets (more generally, \( \left( Q, \leq_Q \right) \) can be any preordered set).

An order homomorphism from \( P \) to \( Q \) is injective if and only if it is strict.

Proof.

Proof of sufficiency. Follows from proposition 1227.

Proof of necessity. Let \( f : P \to Q \) be a strict order homomorphism and suppose that \( f(x) = f(y) \) for some \( x \) and \( y \) in \( P \). We will use the trichotomy of \( <_P \).

- If \( x <_P y \), then \( f(x) <_Q f(y) \) since \( f \) is strictly monotone, which contradicts \( f(x) = f(y) \).
- If \( x >_P y \), similarly \( f(x) >_Q f(y) \) and we again obtain a contradiction.
- It remains for \( x \) to be equal to \( y \).

Since \( x \) and \( y \) were arbitrary, we conclude that \( f \) is injective.

Proposition 1243. For totally ordered sets, strict order homomorphisms are precisely the strong order homomorphisms.

Let \( \left( P, \leq_P \right) \) and \( \left( Q, \leq_Q \right) \) be totally ordered sets (more generally, \( \left( Q, \leq_Q \right) \) can be any preordered set) and let \( f : P \to Q \) be an order homomorphism between them. Then \( f \) is a strict order homomorphism if and only if it is a strong order homomorphism.

Proof.

Proof of sufficiency. Suppose that \( f \) is a strict order homomorphism. We must show that \( f(x) \leq f(y) \) entails \( x \leq y \).

If we suppose that \( x > y \) for some \( x \) and \( y \) in \( P \), then \( f(x) > f(y) \), which contradicts our assumption \( f(x) \leq f(y) \).

Therefore, \( f \) is a strong order homomorphism.

Proof of necessity. Suppose that \( f \) is a strong order homomorphism. Let \( x < y \). It follows that \( f(x) \leq f(y) \).

Suppose that \( f(x) = f(y) \). Then both \( f(x) \leq f(y) \) and \( f(x) \geq f(y) \) hold, which in turn imply that both \( x \leq y \) and \( x \geq y \) hold since \( f \) is a strong homomorphism. Thus, we obtain \( x = y \), which contradicts our assumption that \( x < y \).

Therefore, \( f \) is a strict order homomorphism.
**Corollary 1244.** A strict order embedding between totally ordered sets is an isomorphism if and only if it is bijective.

*Proof.* Follows from proposition 1243 and proposition 861.

**Proposition 1245.** If $(P, \leq)$ is a totally ordered set and $A \subseteq P$ is nonempty, then any minimal element is a minimum.

*Proof.* Let $x_0$ be a minimal element of $A$. If $x_0$ is the only element of $A$, it is clearly the minimum of $A$. Suppose that $A$ is not a singleton set.

By definition of total order, for any $x \in A$ either $x \leq x_0$ or $x_0 \leq x$. If $x \leq x_0$, then since $x_0$ is a minimal element, we have $x = x_0$.

Therefore, for any $x \in A$, either $x = x_0$ or $x > x_0$ (i.e. $x \leq x_0$), proving that $x_0$ is a minimum of $A$.

**Proposition 1246.** Let $(P, \leq)$ be a totally-ordered set. Let $Q$ be the set containing the strict initial segment $P_{<x}$ for every member $x$ of $P$.

Then $(P, \leq)$ is strictly order-isomorphic to $(Q, \subseteq)$.

*Proof.* Explicitly define the isomorphism

$$f : P \rightarrow Q$$

$$f(x) := P_{<x} = \{y \in P \mid y < x\}.$$  

We will first show that $f$ is strictly monotone. If $x < y$, then $x \in P_{<y}$. But $x \notin P_{<x}$, hence $P_{<x}$ is a strict subset of $P_{<y}$. Thus, $f$ is strictly monotone.

The function $f$ is injective by proposition 1242 and surjective by definition. Thus, $f$ is a strict isomorphism between $(P, \leq)$ and $(Q, \subseteq)$.

**Proposition 1247.** Let $(P, \leq)$ be an unbounded from above totally ordered set and let $A \subseteq P$.

Then $A$ is cofinal if and only if it is itself unbounded from above.

This equivalence is useful for regular cardinals — for example proposition 1096.

Compare this result with proposition 1233.

*Proof.*

**Proof of sufficiency.** Let $A$ be a cofinal set. Suppose that it is bounded from above. Then there must exist some $x \in P$ such that $x$ is an upper bound of $A$.

- If $x \notin A$, this contradicts the cofinality of $A$.

- If $x \in A$, then it is a maximum. But since $P$ is itself unbounded from above, we can find some $y \in P$ such that $y$ is again an upper bound of $A$ and $x < y$. But this contradicts the cofinality of $A$.

**Proof of necessity.** Let $A$ be unbounded from above. Suppose that is is not cofinal. Since there exists no upper bound of $A$, for every $x \in P$ there exists some $y \in A$ such that $x \leq y$. Therefore, $A$ is cofinal.
Proposition 1248. If \((P, \leq_P)\) and \((Q, \leq_Q)\) are totally ordered sets, then the lexicographic and reverse lexicographic orders on \(P \times Q\) are strict total order relations.

Compare this result to proposition 1236 and proposition 1001.

Proof. We have already shown in proposition 1236 and these are partial orders. It only remains to check trichotomy.

Proof of trichotomy. Let \(<\) be the lexicographic order on \(P \times Q\). Let \((a, b)\) and \((c, d)\) be pairs in \(P \times Q\). Since \(<_P\) and \(<_Q\) are strict total orders, we only have the following possibilities:

- If \(a = c\) and \(b = d\), then \((a, b) = (c, d)\).
- If \(a = c\) and \(b <_Q d\), then \((a, b) < (c, d)\).
- If \(a <_P c\) and \(b >_Q d\), then \((a, b) > (c, d)\).
- If \(a >_P c\), then \((a, b) < (c, d)\).
- If \(a >_P c\), then \((a, b) > (c, d)\).

The proof for the reverse lexicographic order is analogous. \(\square\)

Definition 1249. Let \(P\) be a totally ordered set with more than one element. The order topology induced by \(\leq\) is the topology generated by the subbase of open rays

\[ S := \{(a, \infty) \mid a \in P\} \cup \{(-\infty, b) \mid b \in P\}. \]

The base corresponding to this subbase is

\[ B = S \cup \{\emptyset\} \cup \{(a, b) \mid a, b \in P \text{ and } a < b\}. \]

See the proof of B1 for why \(P\) must have more than one element.

Proof of correctness.

Proof of compatibility of \(S\) and \(B\). Define

\[ \mathcal{C} = \{\bigcap S \mid S \text{ is a nonempty finite subset of } S\}. \]

We will show that \(B \subseteq \mathcal{C}\).

Let \(B \in \mathcal{B}\). The cases \(B \in S\) and \(B = \emptyset\) are trivial. Suppose that \(B \notin S\). Then there exist points \(a < b\) such that

\[ B = (a, b) = (-\infty, b) \cap (a, \infty). \]

This is an intersection of members of \(S\), hence \(B \in \mathcal{C}\). Therefore, \(B \subseteq \mathcal{C}\).

Now let \(C = S_1 \cap \cdots \cap S_n\), where \(S_1, \ldots, S_n\) are members of \(S\). We will show by induction on \(n > 0\) that \(C \in \mathcal{B}\). The case \(n = 1\) is trivial. Suppose that all \(n\)-ary intersections belong to \(B\) and let

\[ C = S_1 \cap \cdots \cap S_n \cap S_{n+1}. \]

By the inductive hypothesis we have that \(D := S_1 \cap \cdots \cap S_n\) belongs to \(B\) and thus we have three cases:
• If either $D = (a, \infty)$ or $D = (-\infty, b)$, then $D \in S$.

• If $D = \emptyset$, then $D = (-\infty, a) \cap (a, \infty)$ for some $a \in P$.

• If $D = (a, b)$, then $D = (-\infty, b) \cap (a, \infty)$.

In all cases both cases $D$ and $C = D \cap S_{n+1}$ are finite intersection of members of $S$. Therefore, $C \subseteq \mathcal{B}$. Since we already have the inclusion in the other direction, we conclude that $\mathcal{C} = \mathcal{B}$.

**Proof that $\mathcal{B}$ is a base.** We will show that the axioms in proposition 248 hold.

**Proof of B1.** Let $x \in P$.

If $x$ is a **maximum**, then take any other value $y < x$ and the set $(y, \infty)$ will contain $x$. We use here that there is more than one element in $P$.

If $x$ is not a maximum, then $x$ belongs to any interval $(-\infty, y)$ whenever $y > x$.

In both cases there exists an interval in $S$ containing $x$. Thus, $\bigcup S = P$.

**Proof of B2.** Let $U$ and $V$ be members of $\mathcal{B}$. We consider 14 cases:

• If either $U = \emptyset$ or $V = \emptyset$, then $U \cap V = \emptyset$.

• If $U = (-\infty, u)$ and $V = (v, \infty)$, then
  - If $u \leq v$, then $U \cap V = \emptyset$.
  - If $v < u$, then $U \cap V = (v, u)$.

• If $U = (-\infty, u)$ and $V = (v_1, v_2)$, then
  - If $v_1 < v_2 \leq u$, then $U \cap V = (v_1, v_2)$.
  - If $v_1 \leq u < v_2$, then $U \cap V = (v_1, u)$.

• If $U = (u_1, u_2)$ and $V = (v, \infty)$, then
  - If $u_2 \leq v$, then $U \cap V = \emptyset$.
  - If $u_1 \leq v < u_2$, then $U \cap V = (v, u_2)$.
  - If $v \leq u_1 < u_2$, then $U \cap V = U = (u_1, v_1)$.

• If $U = (u_1, u_2)$ and $V = (v_1, v_2)$, then
  - If $u_2 < v_1$, then $U \cap V = \emptyset$.
  - If $u_1 < v_1 < u_2 < v_2$ then $U \cap V = (v_1, u_2)$.
  - If $u_1 < v_1 < v_2 < u_2$ then $U \cap V = (v_1, v_2)$.
  - If $v_1 < u_1 < u_2 < v_2$ then $U \cap V = U = (u_1, u_2)$.
  - If $v_1 < u_1 < v_2 < u_2$ then $U \cap V = (u_1, v_2)$.

In all cases, the intersection $U \cap V$ belongs to $\mathcal{B}$. 

**Example 1250.** Examples of order topologies include:

653
• The order topology on $\mathbb{R}$, which is equivalent to the metric topology as shown in theorem 45.

• All ordinals greater than one induce topological spaces called the ordinal spaces.

**Definition 1251.** Let $\alpha$ be an ordinal. When regarded as the set of smaller ordinals, as shown valid in proposition 1018, $\alpha$ is a totally order set and hence we can endow it with the order topology $\mathcal{T}$ to obtain a topological space. We call the space $(\alpha, \mathcal{T})$ an ordinal space.

**Proposition 1252.** Fix an ordinal space $(\alpha, \mathcal{T})$.

An nonzero ordinal $\beta \in \alpha$ is a limit ordinal if and only if it is the limit point of some net of ordinals in the space $\alpha$.

*Proof.*

**Proof of sufficiency.** Let $\beta$ be a limit ordinal. When regarded as a subset of $\alpha$, it is itself a topological net because it is totally ordered. We will show that $\beta$ as a member of $\alpha$ is a limit point as a subset of $\alpha$.

By proposition 282 it is enough to show that $\beta$ as a net is eventually in every set of the local subbase at $\beta$ of the order topology. This local subbase consists of all initial and final segments of $\alpha$ that contain $\beta$.

With the following we exhaust the local subbase at $\beta$: for all $\gamma \in \alpha$ that are distinct from $\beta$:

• If $\gamma > \beta$, then $\gamma$ itself is a neighborhood of $\beta$ and a member of the subbase as an initial segment of $\alpha$. Since the entire net $\beta$ is contained in $\gamma$, it is eventually in the initial segment $\alpha_{>\gamma}$.

• If $\gamma < \beta$, then the final segment $\alpha_{>\gamma}$ is a member of the local subbase of $\beta$. Let $\delta$ be some member of the net $\beta$.
  
  – If $\delta > \gamma$, it is itself an ordinal such that $\epsilon > \delta$ implies $\epsilon \in \alpha_{>\gamma}$.
  
  – If $\delta \leq \gamma$, then $\epsilon > \text{succ}(\gamma)$ implies $\epsilon \in \alpha_{>\gamma}$. The successor of $\gamma$ belongs to $\beta$ because $\beta$ is a limit ordinal and satisfies definition 1023 (b).

Thus, again the net $\beta$ is eventually in the final segment $\alpha_{>\gamma}$.

We have shown that $\beta$ as a net is eventually in some every set in the local subbase of $\beta$, thus it is a limit of the net.

**Proof of necessity.** Let $\beta$ be a limit of the net $\{\gamma_k\}_{k \in \mathcal{K}} \subseteq \alpha$.

Let $\delta \in \beta$. Consider the neighborhood $(\delta, \beta)$ of $\beta$. Since $\beta$ is a limit point, there must exist some index $k_0 \in \mathcal{K}$ such that $\gamma_k \in \alpha_{>\delta}$ for every $k \geq k_0$. Since $\delta$ was an arbitrary member of $\beta$, we conclude that $\beta$ it satisfies definition 1023 (b) and is thus a limit ordinal. $\Box$
15.4. Lattices

**Definition 1253.** Lattices are partially ordered sets in which suprema and infima are taken as basic operations called “joins” and “meets”. See remark 1254 for a discussion of the operation names. This shifts the focus from ordering to operations, i.e. from predicates to functions.

Joins and meets may also be defined axiomatically as binary operations rather than via some partial order, however this restricts us to taking suprema of finite sets and prevents us from taking the supremum of an arbitrary set. In other words, it is possible for the order to carry more information than joins and meets. See proposition 1255 (e) for a discussion. Unless explicitly noted otherwise, we assume that lattices have their partial order defined.

(a) A **join-semilattice** is a bounded from above partially ordered set in which every finite supremum exists. The operation itself is denoted by \( \lor \) and referred to as **join** and rather than supremum. In contrast to suprema, joins are usually written in infix notation, e.g. \( x \lor y \lor z \) rather than \( \text{sup}\{x, y, z\} \).

(b) Analogously, a **meet-semilattice** is a partially ordered set in which every finite infimum exists. The infimum is denoted by \( \land \) and called **meet**.

(c) A **bounded semilattice** is a semilattice that is bounded as a partially ordered set, either from below for join-semilattices or from above for meet-semilattices.

(d) A semilattice is said to be **complete** if the corresponding operation is defined for arbitrary sets rather than only finite ones.

Finite semilattices are clearly complete, as well as bounded semilattices.

(e) A **lattice** is a partially ordered set which is both a join-semilattice and a meet-semilattice. It is called **bounded** if both semilattices are bounded, i.e. if the partially ordered set itself is bounded. It is called **complete** if both semilattices are complete.

(f) A lattice is said to be **distributive** if the following two conditions hold:

\[
x \lor (y \land z) = (x \lor y) \land (x \lor z)
\]

\[
x \land (y \lor z) = (x \land y) \lor (x \land z).
\]

If the lattice is complete, the above conditions are not enough. A complete lattice \( X \) is said to be **distributive** if any of the following more general distributive axioms hold for every \( x \in X \) and family \( \{y_k\}_{k \in \mathcal{X}} \subseteq X \):

\[
x \lor \left( \bigwedge_{k \in \mathcal{X}} y_k \right) = \bigwedge_{k \in \mathcal{X}} (x \lor y_k)
\]

\[
x \land \left( \bigvee_{k \in \mathcal{X}} y_k \right) = \bigvee_{k \in \mathcal{X}} (x \land y_k)
\]

Lattices have the following metamathematical properties:
(g) The language of the theory of lattices consists of the language of the theory of partially ordered sets with the addition of the binary infix functional symbols \( \lor \) and \( \land \). If we only want to restrict ourselves to semilattices, we can add only one of the two operations as functional symbols. If we wish to study bounded lattices, as it is often done, we must also add the constants \( \top \) and \( \bot \).

For meet-semilattices, we add the following axiom schema to the theory to ensure compatibility between infima and meets (we use \& to denote logical conjunction to avoid symbol collision with meets):

\[
\left( \xi \land \eta \equiv \alpha \right) \leftrightarrow \left( \alpha \leq \xi & \alpha \leq \eta & \forall \alpha . ((\alpha \leq \xi & \alpha \leq \eta) \rightarrow \alpha \leq \alpha) \right)
\] (508)

and, for bounded meet-semilattices, the following axiom to ensure that \( \top \) is indeed the maximum:

\[
\forall \xi . (\xi \leq \top).
\] (509)

Analogous axioms need to be added for join-semilattices.

We cannot properly express the theory of complete (semi)lattices as an extension of this theory since we must define join and meet as unary operations on subsets of the domain rather than binary operations on members of the domain. Complete semilattices can instead be defined within ZFC.

(h) Unlike for partially ordered sets, whose submodels are discussed in definition 1222 (e), not every subset of a semilattice is a sub-semilattice because \( \lor \) and \( \land \) are now regarded as functional symbols. A sub-(semi)lattice must be closed under joins and meets. The axiom (508) is not a positive formula, but does not cause trouble itself as it merely specifies compatibility of \( \leq \) and \( \land \).

For bounded semilattices, the relevant constants should be present in any bounded sub-semilattice.

(i) The trivial join-semilattice and the trivial meet-semilattice are the empty set. The trivial bounded join-semilattice is the singleton \( \{ \bot \} \) and the trivial bounded meet-semilattice is \( \{ \top \} \).

The trivial bounded lattice satisfies \( \top = \bot \), which implies that it consists of one element.

Note that the elements \( \top \) and \( \bot \) formally differ between different semilattices, however all trivial bounded lattices are isomorphic and hence it makes sense to speak of “the” bounded lattice.

(j) Homomorphisms between (semi)lattices are the monotone maps that preserve joins, meets and constants.

As we shall see in corollary 1256, the requirement of monotonicity is redundant.

(k) The categories of \( \mathcal{U} \)-small models for (semi)lattices are full subcategories of \( \mathcal{U} \text{-Pos} \). We only give a special name for the category \( \mathcal{U} \text{-Lat} \) of lattices.
The principle of duality for partially ordered sets holds for lattices if we also swap the binary operations $\vee$ and $\wedge$.

If the lattice is bounded, we must additionally swap the constants $\top$ and $\bot$.

If the lattice is bounded from only one side, the principle of duality does not hold unless we restrict ourselves to formulas that do not contain the constants.

**Remark 1254.** The terms “join” for $\vee$ and “meet” for $\wedge$ are notoriously difficult to remember. A helpful accident is the ability to write “meet” as “$\wedge\wedge$”.

**Proposition 1255.** Let $(P, \leq)$ be a partially ordered set.

(a) If it is a join-semilattice (resp. meet-semilattice), then $\vee$ (resp. $\wedge$) is associative, commutative and idempotent when considered as a binary operation.

(b) If $P$ is a (semi)lattice, the constants act as monoid identities. That is, for each $x \in P$,

\[
x \vee \bot = x \tag{510}
\]
\[
x \wedge \top = x \tag{511}
\]

(c) If $P$ is a lattice, then the following absorption laws hold:

\[
x \vee (x \wedge y) = x \tag{512}
\]
\[
x \wedge (x \vee y) = x. \tag{513}
\]

(d) The following conditions for compatibility with $\leq$ hold:

\[
x \leq y \text{ if and only if } x \vee y = y \tag{514}
\]
\[
x \leq y \text{ if and only if } x \wedge y = x. \tag{515}
\]

(e) If $A$ is an arbitrary set and if $\vee$ is a binary operation that is associative, commutative and idempotent (the conclusion of proposition 1255 (a)), then $(A, \leq)$ is a join-semilattice with an ordering defined by (514). If there exists a distinguished element $\bot$ such that (511) holds, then $(A, \leq)$ is bounded.
A completely analogous statement holds for meet-semilattices.

If \((A, \leq)\) is both a join-semilattice and meet-semilattice and if \(\vee\) and \(\wedge\) satisfy the absorption conditions (512) and (513), then \((A, \leq)\) is a lattice. Furthermore, proving idempotence for \(\vee\) or \(\wedge\) is unnecessary because both follow from the absorption conditions.

It may turn out that \((A, \leq)\) is a complete lattice under this definition. This can allow us, for example, to transparently extend the binary operations join and meet into infinitary operations.

Proof.

**Proof of 1255 (a).** Suprema and infima are obviously associative and commutative as binary operations because ordering is immaterial for pure sets and \(x \lor y\) is defined as \(\text{sup}\{x, y\}\).

Idempotence is also obvious because \(x \lor x = \text{sup}\{x\} = x\).

**Proof of 1255 (b).** Obvious since \(\bot \leq x \leq \top\) for all \(x \in P\).

**Proof of 1255 (c).** If we rewrite (512) using suprema and infima, we obtain

\[
\text{sup}\{x, \text{inf}\{y, x\}\} = x.
\]

If \(x \leq y\), then \(\text{inf}\{y, x\} = x\) and \(\text{sup}\{x, \text{inf}\{y, x\}\} = \text{sup}\{x, x\} = x\).

If \(x \geq y\), then \(\text{inf}\{y, x\} = y\) and \(\text{sup}\{x, \text{inf}\{y, x\}\} = \text{sup}\{x, y\} = x\).

This proves (512). Since \(\wedge\) is \(\lor\) in the opposite partially ordered set, (513) follows automatically.

**Proof of 1255 (d).** We have

\[
x \lor y = \text{sup}\{x, y\} = \begin{cases} y, & x \leq y \\ x, & x > y \end{cases}
\]

and dually for \(\land\).

**Proof of 1255 (e).** Since the binary join and/or meet are defined for all members of the set \(A\), it is indeed a join-semilattice because all finite joins and meets exist by definition.

Idempotence of \(\lor\) follows from (513):

\[
x \lor x = x \lor (x \land (x \lor x)) = x
\]

and dually for \(\land\).

\[\square\]

**Corollary 1256.** If a function \(f : L \to M\) between (semi)lattices preserves either joins or meets, it is monotone.

Thus, the requirement for lattice homomorphisms to be monotone is redundant.

**Proof.** If the function \(f\) preserves joins and if \(x \leq y\), by (514) we have \(x \lor y = y\) and thus

\[
f(x \lor y) = f(y),
\]

which again by (514) implies \(f(x) \leq f(y)\).

\[\square\]
**Proposition 1257.** In any bounded lattice, \( \bot \) is absorbing with respect to meets and \( \top \) with respect to joins. That is, \( \bot \land x = \bot \) and \( \top \lor x = \top \).

**Proof.** Obvious when the lattice is regarded as a partially ordered set. \( \square \)

**Definition 1258.** Given a function \( f \colon A \to A \) between arbitrary sets, we call \( x \in A \) a fixed point of \( f \) if \( x = f(x) \).

**Theorem 1259** (Knaster-Tarski theorem). The fixed points of a monotone endofunction in a complete lattice form a complete sublattice. In particular, the function has at least one fixed point.

**Proof.** Let \((X, \leq)\) be a complete lattice and let \( \varphi : X \to X \) be a monotone function. Define

\[
L := \{ x \in X : f(x) \leq x \}.
\]

We know that \( L \) is nonempty because \( \top \in L \).

Since the lattice is complete, we can take \( l := \inf L \). Note that \( f(l) \) is a lower bound of \( L \) because for any \( y \in L \) we have

\[
f(l) \leq f(y) \leq y.
\]

But \( l \) is the largest lower bound of \( L \), hence

\[
f(l) \leq l. \tag{516}
\]

Therefore, \( f(f(l)) \leq f(l) \) and \( f(l) \in L \). Hence, \( l \) is a lower bound for \( \{f(l)\} \) and

\[
l \leq f(l). \tag{517}
\]

From (516) and (517) it follows that \( l = f(l) \), that is, \( l \) is a fixed point of \( f \).

Denote by \( F \) the set of all fixed points of \( X \). We just showed that \( F \) is nonempty. Let \( G \subseteq F \).

We will show that the infimum and supremum of \( G \) is in \( F \).

Denote

\[
l_G := \inf G.
\]

For any \( g \in G \) we have \( l_G \leq g \). From monotonicity of \( f \),

\[
f(l_G) \leq f(g) = g,
\]

therefore \( f(l_G) \leq l_G \) because \( l_G \) is the greatest lower bound of \( G \). But, from monotonicity of \( f \), we have \( l_G \leq f(l_G) \). Therefore, \( f(l_G) = l_G \) and \( l_G \in F \).

We can analogously show that \( \sup G \in F \) and conclude that \((F, \leq)\) is itself a complete lattice. \( \square \)

**Remark 1260.** The existence of finite joins and meets is equivalent to the existence of finite products and coproducts in the respective thin category defined in theorem 1186 (b).

**Definition 1261.** An element \( x \) of a commutative semiring is said to be square-free if \( y \mid x \) implies that \( z^2 \nmid x \).
Remark 1262. Let $L$ be a bounded distributive lattice. We will discuss polynomials over $L$.

By example 527 (d), $L$ induces a positive and a negative commutative semiring. Given a set $\mathcal{X}$ of indeterminates, we can form the polynomial semiring $L[\mathcal{X}]$ over the positive semiring. Suppose that we are given an evaluation $f : \mathcal{X} \to L$. Consider the evaluation homomorphism $\Phi_f : L[\mathcal{X}] \to L[\mathcal{X}]$.

$$\Phi_f(X^2) = \Phi_f(X).$$

Hence, we can limit ourselves to square-free monomials. That is, monomials of the form $\prod_{X \in \mathcal{X}} X^{\gamma_X}$, where $\gamma_X$ is either 0 or 1. More succinctly, since every monomial has finitely many indeterminates of positive power, it can be written as a finite meet $X_1 \land \cdots \land X_n$. A polynomial is then a finite join of finite meets of indeterminates and constants.

There is a nuance, however. In [Gra78, def. I.4.2], a multivariate lattice polynomial is defined to consist only of indeterminates; for example $p(X, Y, Z) = (X \land Y) \lor (X \land Z) \lor (Y \land Z)$.

This excludes coefficients before the monomials, hence making the definition distinct from the general notion of a polynomial over a commutative semiring. In [Mar07], polynomials with coefficients in front of the monomials are called weighted lattice polynomials. For example, a weighted polynomial is

$$q(X, Y, Z) = (a \land X \land Y) \lor (b \land X \land Z) \lor (c \land Y \land Z).$$

Compare $q(X, Y, Z)$ to $p(X, Y, Z)$. We will refer to unweighted polynomials by default.

We can analogously define polynomials over the negative semiring of $L$, e.g.

$$r(X, Y, Z) = (X \lor Y) \land (X \lor Z) \land (Y \lor Z).$$

The latter polynomials are related to but distinct from conjunctive normal forms, while the former - to disjunctive normal forms.

**Example 1263.** We list several examples of lattice polynomials:

(a) The distributivity axiom (505) implies that the polynomial $X \land (Y \lor Z)$ over the positive semiring evaluates to the same trivariate function as the polynomial $(X \land Y) \lor (X \land Z)$ over the negative semiring.

(b) A lattice polynomial for the two-element boolean algebra $\{T, F\}$ corresponds to a boolean function. We cannot express negation without introducing auxiliary polynomials as a consequence of example 803 (a). Positive formulas in conjunctive/disjunctive
normal form, however, correspond exactly to lattice polynomials with square-free monomials.

For example, exclusive or $\oplus$ can be expressed via the propositional formula

$$P \lor Q \lor (P \land Q).$$

This corresponds to a bivariate polynomial over the negative semiring of $\{T, F\}$.

**Definition 1264.** In a partially ordered set, we call the nonempty upward directed subset $I$ an order ideal if $x \in I$ and $y \leq x$ imply that $y \in I$.

Dually, we call the nonempty downward directed set $F$ a filter if $x \in F$ and $y \geq x$ imply that $y \in F$.

**Proposition 1265.** For a subset $I$ of a lattice $L$, the following are equivalent:

(a) $I$ is an order ideal.

(b) $I$ is a closed under joins and $i \in I$ and $l \in L$ imply $i \land l \in I$. If $L$ has a bottom, it must belong to $I$.

(c) $I$ is a semiring ideal of the positive semilattice of $L$ (in case $L$ is a bounded distributive lattice).

**Proof.**

**Proof that 1265 (a) implies 1265 (b).** Suppose that $I$ is an order ideal.

- The bottom $\perp$ is less than any element of $I$, hence $\perp \in I$.
- Given $i, j \in I$, $I$ contains an upper bound of theirs. Then $i \lor j$ as the least upper bound also belongs to $I$.
- Given $i \in I$ and $l \in L$, since $i \land l \leq l$, we conclude that $i \land l \in I$.

**Proof that 1265 (b) implies 1265 (a).** Suppose that $I$ is closed under joins and that $i \in I$ and $l \in L$ imply $i \land l \in I$.

- If $i, j \in I$, then $i \lor j \in I$. Hence, $I$ is an upward directed set.
- If $i \in I$ and $j \leq i$, then $j = i \land j \in I$.

**Proof of equivalence of 1265 (a) and 1265 (c).** Trivial. □

**Remark 1266.** Regarding ideals and filters in bounded distributive lattices as semiring ideals exposes us to a lot of definitions and theorems that we would otherwise need to redefine. For example, principal, prime and maximal ideals, the properties from proposition 558, the semiring of ideals from proposition 559, or theorem 561 (Maximal ideal theorem).

**Example 1267.** We list examples of lattice ideals:
Figure 35: A Hasse diagram for the divisors of 24.

(a) Consider the (zero-based) natural number divisibility lattice from proposition 12. For any natural number \( n \), the set \( D_n \) of all divisors of \( n \) is a lattice ideal. Indeed,
- The bottom 1 divides \( n \).
- If \( a \) and \( b \) divide \( n \), their product \( ab \) also does, and hence their join \( \text{lcm}(a, b) \) also does.
- If \( a \mid n \) and \( b \) is any natural number, then \( \gcd(a, b) \mid a \mid n \).

Furthermore, \( D_n \) is a principal ideal since it can be obtained as \( \{ n \wedge m \mid m \in \mathbb{N} \} \).

If \( p \) is a prime number, then \( D_p = \{1, p\} \) is a maximal ideal, and hence also a prime ideal.

(b) Given a group \( G \) and a proper normal subgroup \( N \), consider the lattice of subgroups \( L_G \) of \( G \) and the sublattice \( L_N \) of subgroups of \( N \). Note that the top \( N \) of \( L_N \) is not the top \( G \) of \( L_G \).

Then \( L_N \) is an ideal on \( L \).
- \( L_N \) contains the bottom \( \{e\} \).
- \( L_N \) contains the join \( \langle K \cup H \rangle \) of every two subgroups of \( N \).
- \( L_N \) contains the meet \( K \cap H \) of \( K \in L_N \) and \( H \in L_G \) since \( K \cap H \subseteq K \subseteq N \).

**Proposition 1268.** For a subset \( F \) of a lattice \( L \), the following are equivalent:

(a) \( F \) is a filter.

(b) \( F \) is closed under meets and \( i \in I \) and \( l \in L \) imply \( i \lor l \in I \). If \( L \) has a top, it must belong to \( I \).

(c) \( F \) is a semiring ideal of the negative semilattice of \( L \) (in case \( L \) is a bounded distributive lattice).

**Proof.** The proof is dual to that of proposition 1265. □
Example 1269. We list examples of lattice filters:

(a) By duality with respect to example 1267 (b), it follows that, for a natural number $n$, the semiring ideal

$$\langle n \rangle := \{0, n, 2n, 3n, \ldots\}$$

is a principal filter in the natural number divisibility lattice from proposition 12.

We can prove this explicitly:

- $\langle n \rangle$ contains the top 0.
- $\langle n \rangle$ contains the meet $\gcd(a, b)$ of every two members of $\langle n \rangle$.
- If $n \mid a$ and $b$ is any natural number, then $n \mid a \mid \lcm(a, b)$, meaning that the join $\lcm(a, b)$ also belongs to $\langle n \rangle$.

(b) By duality with respect to example 1267 (b), given a group $G$ and a proper normal subgroup $N$, the sublattice of subgroups containing $N$ (rather than contained in $N$) is a filter (rather than an ideal).
15.5. Boolean algebras

**Definition 1270.** A **Heyting algebra** is a bounded distributive lattice $X$ with a binary operation $\rightarrow$ defined as

$$ (x \rightarrow y) := \bigvee \{a \in X \mid a \land x \leq y\}. \quad (518) $$

In order for the operation to be well-defined, we require that the corresponding join exists for all $y$ and $z$. We call this operation the **conditional** in analogy with the propositional connective, although this operation is often called “implication” because of material implication.

Heyting algebras are useful for defining truth values for intuitionistic logic—see definition 890—and also appear as the Lindenbaum-Tarski algebra for propositional intuitionistic theories—see proposition 919.

(a) For any element $x$, we define its **pseudocomplement** as

$$ \bar{x} := (x \rightarrow \bot) = \bigvee \{a \in X \mid a \land x = \bot\}. \quad (519) $$

Heyting algebras have the following metamathematical properties:

(b) We extend the language of the **theory of lattices** with the binary infix functional symbol $\rightarrow$ and the unary functional symbol $\bar{\cdot}$. By adding the axiom (518) to the theory of bounded distributive lattices, we obtain the theory of Heyting algebras.

(c) The Heyting subalgebras are the **bounded sublattices** for which the conditional is well-defined.

(d) The **trivial Heyting algebra** is the one-element bounded lattice.

(e) **Homomorphisms** between Heyting algebras are lattice homomorphisms with the additional requirement that homomorphisms preserve conditionals.

(f) The **category of $\mathcal{U}$-small models** for Heyting algebras is denoted by $\text{Heyt}$. It is a full subcategory of the category $\text{Lat}$ of lattices.

(g) The **principle of duality for lattices** does not hold for Heyting algebras.

**Example 1271.** Somewhat similar to how the power set of a nonempty set is a Boolean algebra, as shown in proposition 938, the topology $\mathcal{T}$ of a topological space $(X, \mathcal{T})$ is a Heyting algebra. This is actually used in topological semantics—see definition 893.

Indeed,

- Arbitrary joins are given by unions $\bigcap$.
- Finite meets are given by intersections $\bigcup$.
- The top is the entire domain $X$.
- The bottom is the empty set.
The conditional $U \hookrightarrow V$ is then
\[
\bigcup\{A \in \mathcal{T} \mid A \cap U \subseteq V\} = \bigcup\{A \in \mathcal{T} \mid A \subseteq V \cup (X \setminus U)\} = \text{int}((X \setminus U) \cup V),
\]
which is actually similar to proposition 815 (d) despite the fact that arbitrary topologies are not Boolean algebras.

As a result, the pseudocomplement is
\[
\bar{U} = \text{int}(X \setminus U).
\]

**Definition 1272.** Let $X$ be a bounded lattice and fix an element $x \in X$. A complement of $x$ is an element $y$ such that
\[
x \lor y = \top, \quad (520)
\]
\[
x \land y = \bot. \quad (521)
\]

Due to the commutativity of both $\lor$ and $\land$, $y$ is a complement of $x$ if and only if $x$ is a complement of $y$.

**Proposition 1273.** In a bounded distributive lattice $X$, each $x \in X$ has at most one complement. Therefore, complementation can be regarded as a partial operation.

**Proof.** If $y$ and $z$ are both complements of $x$, then
\[
y \overset{(511)}{=} y \land \top =
\]
\[
\overset{(520)}{=} y \land (z \lor x) =
\]
\[
\overset{(503)}{=} (y \land z) \lor (y \land x) =
\]
\[
\overset{(521)}{=} y \land z =
\]
\[
\overset{(512)}{=} (x \land z) \lor (y \land z) =
\]
\[
\overset{(503)}{=} (x \lor y) \land z =
\]
\[
\overset{(513)}{=} z.
\]

**Definition 1274.** A Boolean algebra is a bounded distributive lattice in which every element has a complement. The complement of each element is unique due to proposition 1273. We define a unary function that gives to every element $x$ its complement $\bar{x}$. By definition, this function is an involution.

(a) We also define the binary operation conditional $(\rightarrow)$ via
\[
(x \rightarrow y) := (\bar{x} \lor y) \quad (522)
\]
in analogy with proposition 815 (d). This operation highlights that Boolean algebras are a special case of Heyting algebras.

[106x747]∙
(b) It remains to define a binary operation corresponding to the propositional biconditional. Inspired by proposition 815 (e), define

\[(x \leftrightarrow y) := (x \to y) \land (y \to x).\]  \hspace{1cm} (523)

Boolean algebras have the following metamathematical properties:

(c) To obtain the theory of Boolean algebras, we replace the unary functional symbol \(\neg\) with \(\bar{x}\) in the language of the theory of Heyting algebras and then add the axioms (520) and (521) to the theory. We may also replace (518) defining \(\to\) with the simpler axiom (522).

(d) The Boolean subalgebras are the bounded sublattices which are closed under complementation.

(e) The trivial Boolean algebra is the one-element bounded lattice.

(f) Homomorphisms between Boolean algebras are simply lattice homomorphisms. Complements are automatically preserved because for any lattice homomorphism \(\varphi\) between the Boolean algebras \(X\) and \(Y\),

\[\varphi(x) \lor_Y \varphi(\overline{x}) = \varphi(x \lor_X \overline{x}) = \varphi(T_X) = T_Y,\]

and similarly for \(\land\), hence, due to proposition 1273,

\[\varphi(\overline{x}) = \overline{\varphi(x)}.\]

Implications are also automatically preserved because of (522).

(g) The category of \(\mathcal{U}\)-small models for Boolean algebras \(\mathbf{Bool}\) is a full subcategory the category \(\mathbf{Heyt}\) of Heyting algebras.

(h) The principle of duality for lattices holds for Boolean algebras without interchanging complements.

**Example 1275.** Examples of Boolean algebras include:

- The Lindenbaum-Tarski algebras for classical logic. We prove in proposition 919 that it is a Boolean algebra.
- The prime field \(\mathbb{F}_2\) with suitably defined operations discussed in corollary 1277.
- The power set of any set, usually taken to be a space with additional structure (see proposition 938).

**Proposition 1276.** All two-element Boolean algebras are isomorphic.

*Proof.* Follows from the requirement that lattice homomorphisms preserve constants. \(\square\)
**Corollary 1277.** For certain purposes, for example Zhegalkin polynomials, we may regard the prime field $\mathbb{F}_2$ as a Boolean algebra with 1 as the top and 0 as the bottom element.

**Proof.** Follows from proposition 1276. \qed

**Proposition 1278.** Every Boolean algebra is a Heyting algebra with an identification given by (522).

**Proof.** Fix any $x, y \in X$ in a Boolean algebra $X$. We will show that $x \to y$ as defined in (522) satisfies (518).

Let

$$A := \{a \in X \mid a \land x \leq y\}$$

be the set from (518).

We will show that $\overline{x} \lor y$ is an upper bound of $A$.

Fix some $a_0 \in A$. By definition of $A$, we have

$$a_0 \land x \leq y.$$ 

But this means that

$$\frac{(a_0 \land x) \lor \overline{x}}{a_0 \lor \overline{x}} \leq y \lor \overline{x}.$$ 

Since $a_0 \leq a_0 \lor b$ for any $b \in X$, it follows that $a_0 \leq y \lor \overline{x}$. Therefore, $\overline{x} \lor y$ is indeed an upper bound of $A$.

Also note that

$$(\overline{x} \lor y) \land x = (\overline{x} \lor x) \land (y \land x) = y \land x \leq y,$$

hence $\overline{x} \lor y \in A$.

Thus, $\overline{x} \lor y$ is both an upper bound of $A$ and an element of $A$, i.e. it is the least upper bound of $A$. Therefore,

$$\overline{x} \lor y = \bigvee A.$$

\qed

**Theorem 1279** (De Morgan’s laws). If $X$ is a Boolean algebra, the following hold for any finite family $\{x_k\}_{k \in \mathcal{K}} \subseteq X$:

$$\bigvee_{k \in \mathcal{K}} x_k = \bigwedge_{k \in \mathcal{K}} \overline{x_k} \quad (524)$$

$$\bigwedge_{k \in \mathcal{K}} x_k = \bigvee_{k \in \mathcal{K}} \overline{x_k} \quad (525)$$

If $X$ is complete, $\mathcal{K}$ may be taken to be any family, not necessarily finite.
Proof. We will only show (524) since (525) is dual.

In order for \( \bigwedge_{k \in \mathcal{K}} x_k \) to be the complement of \( \bigvee_{k \in \mathcal{K}} x_k \), the conditions (520) and (521) need to be satisfied.

From distributivity, we have

\[
\left( \bigwedge_{k \in \mathcal{K}} x_k \right) \lor \left( \bigvee_{m \in \mathcal{K}} x_m \right) \overset{(506)}{=} \bigwedge_{k \in \mathcal{K}} \left( x_k \lor \bigvee_{m \in \mathcal{K}} x_m \right) = \bigwedge_{k \in \mathcal{K}} \left( x_k \lor \bigvee_{m \in \mathcal{K} \setminus \{k\}} x_m \right) = \bigwedge_{k \in \mathcal{K}} \top = \top,
\]

which proves (520). The proof of (521) is analogous. \( \square \)
16. Combinatorics

Combinatorics originated as the study of problems related to counting, although this study is now referred to as enumerated combinatorics to distinguish it from more abstract subfields. A canonical example of a counting problem is given in example 1280. On the other hand, proposition 1290 provides an example of how mathematical analysis can answer questions related to counting. The latter subfield is called analytic combinatorics.

Results in combinatorics are traditionally obtained for finite sets and positive integers, however many of them can be easily generalized. For example, theorem 1284 (Dirichlet’s pigeonhole principle) is stated in terms of cardinals rather than positive integers and graphs are not necessary finite.

Some basic definitions and theorems are stated in section 16.1 (Enumerative combinatorics) and section 16.2 (Progressions), however most of the section is concerned with graphs — see section 16.3 (Hypergraphs), ?? ([UNDEFINED]), section 16.6 (Trees).

As usual, in this section $\mathbb{K}$ will refer to either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers. This restriction is justified by remark 401.

Example 1280. A simple, but nontrivial counting problem is due to Fibonacci. In [OR98], the problem is formulated as follows:

A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?

The rabbit behavior described in ridiculously idealized. Pregnancy time, for example, seems to be nonexistent in this problem. We are interested in counting, however, and not in biology. Nonetheless, the problem as it is stated is still open to interpretations. To obtain Fibonacci’s result, we must add some further assumptions:

- The goal is to find the cumulative rabbit count, including the first pair.
- No crossbreeding is assumed — each rabbit is monogamous and is part of a predefined pair.
- The rabbits do not die, which is a realistic assumption for a short time period in case the rabbits are taken care of.
- During the first two months, no rabbit is born. That is, the original pair must also wait for two months prior to producing offspring.

Instead of trying to give a direct answer to the problem, the usual approach is to iteratively build a sequence $\{F_k\}_{k=1}^\infty$, the Fibonacci sequence, indicating the cumulative number of pairs of rabbits each month.

As mentioned, during the first two months no offspring is produced, so we only have one pair of rabbits. Thus, $F_1 = F_2 = 1$. In general, the number of rabbits on month $n$ is the sum of:
• The number of existing pairs on month \( n - 1 \), which is \( F_{n-1} \).

• The number of newborn pairs, which is precisely the number of mature pairs (those born at least two months ago). This is \( F_{n-2} \).

The entire sequence can thus be built using the following recursive definition:

\[
F_k := \begin{cases} 
1, & k = 1 \text{ or } k = 2, \\
F_{k-1} + F_{k-2} & k > 2.
\end{cases}
\]

Just to give an answer to Fibonacci’s question, we will list the first 12 Fibonacci numbers:

\[1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233\]

It should also be noted that outside the rabbit problem, the Fibonacci sequence is often defined to start at the zeroth month with a value of zero — this is how it is defined in remark 982 actually. So

\[
F_k := \begin{cases} 
0, & k = 0, \\
1, & k = 1, \\
F_{k-1} + F_{k-2} & k > 2.
\end{cases}
\]

is actually a more conventional definition.

**Definition 1281.** Let \( A \) be an arbitrary set in the sense of ZFC. Suppose that for every member \( x \in A \) we are given a weight \( w_x \). This weight may be any other set (in ZFC).

The function \( W(x) := w_x \) is called a **weighted set**. It is a function by the axiom schema of replacement.

We call the set \( A \) the **universe** of \( W \). If \( x \in A \), we say that \( x \) belongs to \( W \) and denote this by \( x \in W \), although this is not actual set membership. We denote the weight of \( x \) using \( W(x) \).

Weighted sets are also called **labeled sets**, in which case \( W(x) \) is called the **label** of \( x \) rather than the weight.

We list several important special cases:

(a) If the weights are cardinal numbers, \( W \) is called a **multiset** and the weights are called **multiplicities**. The **multiset cardinality** of \( W \) is the sum of all multiplicities.

(b) If the weights are real numbers in \( [0, 1] \), \( W \) is called a **fuzzy set**.

**Example 1282.** We list several examples of weighted sets.

• If \( G = (V, E) \) is a directed graph, either \( V \) or \( E \) may be a weighted set. Both occur frequently in applications.

• The roots of any polynomial form a multiset with multiplicities given by ?? ([UNDEFINED]).

• Any factorization in an integral domain produces a multiset.
• The point spectrum of a linear operator produces a multiset.

*Remark* 1283. If it is clear from the context that $M$ is a *multiset*, we may write

$$M = \{a, a, b, b, c\}$$

to denote the function

$$M = \{(a, 2), (b, 2), (c, 1)\}.$$  

Since the ordering of elements is irrelevant, we can also regard a multiset as an equivalence class of *transfinite sequences*. 
16.1. Enumerative combinatorics

This subsection lists several results of various importance that don’t really belong to any more consistent theory.

**Theorem 1284** (Dirichlet’s pigeonhole principle). *If we are given more pigeons than pigeonholes, then at least one pigeonhole must contain multiple pigeons in it.*

More formally, if \( \text{card}(A) > \text{card}(B) \), then there exists no injective function from \( A \) to \( B \).

*Proof.* This is a corollary of corollary 1043. \( \square \)

**Definition 1285.** The *binomial coefficient* of the nonnegative integers \( n \) and \( k \) is

\[
\binom{n}{k} := \frac{n!}{k!(n-k)!}
\]

They are motivated by **Theorem 1287** (Newton’s binomial theorem).

**Theorem 1286** (Pascal’s identity). Binomial coefficients have the following property:

\[
\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.
\] (527)

*Proof.*

\[
\binom{n-1}{k} + \binom{n-1}{k-1} = \frac{(n-1)!}{k!(n-1-k)!} + \frac{(n-1)!}{(k-1)!(n-1-k)!} = \frac{(n-1)!}{(k-1)!(n-1-k)!} \left[ \frac{1}{k} + \frac{1}{n-k} \right] = \frac{(n-1)!}{(k-1)!(n-1-k)!} \frac{n}{k(n-k)} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.
\]

\( \square \)

**Theorem 1287** (Newton’s binomial theorem). *If, in some semiring, the members \( x \) and \( y \) commute (i.e. \( xy = yx \)), then*

\[
(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.
\] (528)

*Proof.* We use induction on \( n \). For \( n = 0 \), the theorem trivially holds. Assume that the theorem holds for \( 1, \ldots, n \). Then

\[
(x + y)^{n+1} = x(x + y)^n + y(x + y)^n =
\]
\[
\sum_{k=0}^{n} \binom{n}{k} x^{k+1} y^{n-k} + y \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \\
= x^{n+1} + y \sum_{k=0}^{n-1} \binom{n}{k} x^{k+1} y^{n-(k+1)} + y \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \\
= x^{n+1} + y \left( \sum_{k=1}^{n} \binom{n}{k-1} x^k y^{n-k} + y^n \sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k} \right) + y^{n+1} = \\
\overset{(527)}{=} x^{n+1} + y \sum_{k=1}^{n} \binom{n+1}{k} x^k y^{n-k} + y^{n+1} = \\
= \sum_{k=0}^{n} \binom{n+1}{k} x^k y^{(n+1)-k}.
\]

\[\square\]

**Proposition 1288.** For every ring element \(x\) and every nonnegative integer \(n\), we have

\[x^{n+1} - y^{n+1} = (x - y)(x^n + x^{n-1}y + \cdots + y^n) = (x - y) \sum_{k=0}^{n} x^k y^{n-k}. \tag{529}\]

**Proof.**

\[
(x - y) \sum_{k=0}^{n} x^k y^{n-k} = \sum_{k=0}^{n+1} x^{k+1} y^{n-k} - \sum_{k=0}^{n} x^k y^{n-k+1} = \\
= \sum_{k=0}^{n+1} x^k y^{(n+1)-k} - \sum_{k=0}^{n} x^k y^{(n+1)-k} = x^{n+1} - y^{n+1}.
\]

\[\square\]

**Definition 1289.** The factorial of a nonnegative integer \(n\) is defined recursively as

\[n! := \begin{cases} 1, & n = 0 \\ (n-1)!n, & n > 0. \end{cases}\]

**Proposition 1290.** For every nonnegative integer \(n\) we have

\[\Gamma(n + 1) := n!,\]

where \(\Gamma\) is the Gamma function defined in definition 153.

**Proof.** We use induction on \(n\).

- **If \(n = 0\), then**

\[
\Gamma(1) = \int_{0}^{\infty} x^0 e^{-x} dx = -e^{-x}|_{x=0}^{\infty} = -\lim_{x \to \infty} e^{-x} + 1 = 1 = 0!
\]
If \( n > 0 \) and \( \Gamma(n) = (n - 1)! \), then

\[
\Gamma(n + 1) = \int_0^\infty x^n \cdot e^{-x} \, dx =
\]

\[
= \left. (-x^n e^{-x}) \right|_0^\infty + n \int_0^\infty e^{-x} x^{n-1} \, dx =
\]

\[
= n\Gamma(n) =
\]

\[
= n(n - 1)! =
\]

\[
= n!
\]

\[\square\]

**Theorem 1291** (Stirling’s factorial approximation). *For every nonnegative integer \( n \) there exists some constant \( \theta \in (0, 1) \) such that*

\[
n! = \sqrt{2\pi n} \cdot \left( \frac{n}{e} \right)^n \cdot e^{\frac{1}{12n\theta}}.
\]

*Proof.* Follows from *proposition 1290* and *theorem 154* (Stirling’s gamma approximation). \[\square\]
16.2. Progressions

Progressions are an elementary concept that happens to be useful quite often. There is no
definition of progression, but rather the term “progression” refers to specific recursively
defined sequences.

**Definition 1292.** The **arithmetic progression** with base $a_0$ and difference $d$ is the se-
quence

$$a_k := \begin{cases} a_0, & k = 0, \\ a_{k-1} + d, & k > 0. \end{cases} \quad (530)$$

Clearly every index $k \geq 0$ we have the closed form representation $a_k = a_0 + kd$.

**Proposition 1293.** The series constructed from the arithmetic progression (530) has partial
sums

$$\sum_{k=0}^{n} a_k = \frac{(n + 1)(a_n - a_0)}{2}. \quad (531)$$

In the special case where $a_0 = 0$ and $d = 1$, this reduces to

$$\sum_{k=0}^{n} k = \frac{n(n + 1)}{2}. \quad (532)$$

**Proof.**

$$2 \sum_{k=0}^{n} a_k = 2 \sum_{k=0}^{n} (a_0 + kd) =$$

$$= \sum_{k=0}^{n} (a_0 + kd) + \sum_{k=0}^{n} (a_0 + (n - k)d) =$$

$$= \sum_{k=0}^{n} (2a_0 + nd) =$$

$$= (n + 1)(a_0 + a_n).$$

\hfill \Box

**Definition 1294.** The **geometric progression** with base $a_0$ and denominator $q$ is the se-
quence

$$a_k := \begin{cases} a_0, & k = 0, \\ a_{k-1}q, & k > 0. \end{cases} \quad (533)$$

Clearly every index $k \geq 0$ we have the closed form representation $a_k = a_0q^k$.

The series

$$\sum_{k=0}^{\infty} a_k = a_0 \sum_{k=0}^{\infty} q^k. \quad (534)$$

is called the **geometric series** for $q$. Without loss of generality, we will assume $a_0 = 1$ when
speaking about geometric series.
**Proposition 1295.** The geometric series (534) has the following basic properties:

(a) For all \( q \in \mathbb{C} \setminus \{1\} \), the geometric series (534) has partial sums

\[
\sum_{k=0}^{n} q^k = \frac{1 - q^{n+1}}{1 - q}.
\]

(535)

Compare this to proposition 1288.

(b) In the degenerate case \( q = 1 \), the progression itself is constant, and its partial sums are instead

\[
\sum_{k=0}^{n} q^k = n + 1.
\]

(536)

(c) For \( |q| \geq 1 \), the geometric series diverges.

(d) For \( 0 < |q| < 1 \), the geometric series converges absolutely with sum

\[
\sum_{k=0}^{\infty} q^k = \frac{1}{1 - q}.
\]

(537)

**Proof.**

**Proof of 1295 (a).** Follows from proposition 1288.

**Proof of 1295 (b).** Obvious.

**Proof of 1295 (c).** For \( q = 1 \), proposition 1295 (b) implies that the series diverges because it grows indefinitely. If \( |q| = 1 \) and \( q \neq 1 \), the integer powers \( q^k \) are rotations around the complex plane unit circle, which do not tend to a limit. Hence, the series diverges again.

When \( |q| > 1 \), \( |q^n| \) grows indefinitely with \( n \), and it follows that

\[
\sum_{k=m}^{n} q^k = q^m \sum_{k=0}^{n-m} q^k = q^m \frac{1 - q^{n-m+1}}{1 - q} = \frac{q^m - q^{n+1}}{1 - q}.
\]

can get arbitrarily large. Therefore, in this case the series also diverges.

**Proof of 1295 (d).** Fix \( q \in B(0,1) \). Since only \( q^{n+1} \) depends on \( n \) in (535), we obtain (537) by simply noting that \( q^n \to 0 \) when \( n \to \infty \).

\( \square \)

**Example 1296.** A simple but important practical example of a geometric series is

\[
\sum_{k=0}^{\infty} \frac{1}{2^k} = \frac{1}{1 - 1/2} = 2.
\]

(538)

Note that if the series starts at \( k = 1 \) instead of \( k = 0 \), it sums to 1. This is often applied in analysis indirectly via corollary 119.

Another application of (538) is showing that \( \sum_{k=0}^{\infty} \frac{1}{2^k} = 2 \) in the binary numbers system. More generally, for the \( n \)-ary number system we have

\[
\sum_{k=0}^{\infty} \left( \frac{n-1}{n} \right)^k = \frac{1}{1 - (n-1)/n} = n.
\]

(539)
Remark 1297. In this example we exploit the equivalence between the closed form representations in definition 1292 and definition 1294 and the corresponding inductive definitions. The equivalences are obvious from a mathematical standpoint, however outside of mathematics they have highly nontrivial consequences. Indeed, they highlight the difference between simple interest and compound interest.

As an example, a savings account with $1000 with a simple monthly interest of 2% will earn $240 over a year:

$$1000\left(1 + \frac{2}{100}\right)^{12} = 1240.$$  

The same account with a compound interest of 2% will earn a bit more - about $268:

$$1000\left(1 + \frac{2}{100}\right)^{12} \approx 1268.24.$$  

Over the course of ten years, however, simple interest will earn a total of $2400, while compound interest will earn $\approx 9765$.

The difference between linear and exponential growth appears staggering in a real world situation even though the difference may not be very noticeable short-term.

Definition 1298. The harmonic progression with base $a_0$ and difference $d$ is the sequence

$$a_k := \frac{1}{a_0 + kd}.$$  \hfill (540)

That is, each term is the reciprocal of the corresponding term in an arithmetic progression with the same base and difference. In order for (540) to be well-defined, either

- $d = 0$ and $a_0 \neq 0$, which turns (540) into the constant sequence $\{1/a_0\}_{k=0}^\infty$.
- $d \neq 0$, in which case

$$a_k = \frac{d}{a_0/d + k}.$$  

Thus, if $d \neq 0$, $a_0/d$ must not be a negative integer unless we are satisfied with only the first $-a_0/d$ terms of the progression existing.

Furthermore, the series may only start at $k = 0$ if $a_0 \neq 0$.
For series related to harmonic progressions, see example 125

Remark 1299. Unlike definition 1292 and definition 1294, we have defined the harmonic progressions via closed-form expressions. Indeed, the equivalent inductive definition is more awkward to work with:

$$a_k := \begin{cases} 
1/a_0, & k = 0, \text{ only defined if } a_0 \neq 0, \\
1/(a_0 + d), & k = 1, \\
1/(1/a_{k-1} + d), & k > 0.
\end{cases}$$
16.3. Hypergraphs

**Definition 1300.** Fix a set $V$, whose members we will call **vertices** or **nodes**, and a disjoint from $V$ set $E$, whose members we will call **hyperedges**.

A **hypergraph** is a tuple $H = (V, E, \mathcal{E})$, where $\mathcal{E} : E \rightarrow V$ is a total multi-valued function whose role is to give us a nonempty set of endpoints $\mathcal{E}(e)$ for each edge $e \in E$.

![Hypergraph Diagram](541)

Figure 36: A hypergraph containing four hyperedges of rank two and one of rank four.

(a) The **cardinality** of a hyperedge $e$ is the cardinality of the set of all vertices of $e$.

Hyperedges of rank 1 are called **loops** and hyperedges of rank 2 are called **edges**. Loops are often considered to be edges, especially in **multigraphs**.

(b) We say that the hyperedges $e$ and $d$ are parallel if they have the same endpoints.

(c) We say that the vertex $v$ and the hyperedge $e$ are **incident** if $v$ is an endpoint $e$.

(d) We say that $v$ and $w$ are **adjacent vertices** if there exists an hyperedge $e$ such that both $v$ and $w$ are endpoints of $e$.

Similarly, we say that $e$ and $d$ are **adjacent hyperedges** if they have a common endpoint.

(e) The **order** $\text{ord}(H)$ of the hypergraph $H$ is the cardinality of $V$.

We say that the hypergraph is finite if both $V$ and $E$ are finite and infinite otherwise. If no parallel hyperedges are allowed, the hypergraph is finite if $\text{ord}(H)$ is finite.

(f) The **degree** $\text{deg}(v)$ of a vertex $v \in V$ is the cardinality of set

\[ \{e \in E \mid v \text{ is an endpoint of } e\}. \]

If $\text{deg}(v) = 0$, we say that $v$ is an **isolated vertex**, especially in connection with definition 1342 (a).

The degree $\text{deg}(H)$ of the hypergraph itself is the maximum of the degrees of all vertices. It is possible that the maximum in $\text{deg}(H)$ is not attained if $H$ is infinite.
We say that the graph is **locally finite** if the degree of every vertex is finite. This is not to be confused with **locally finite categories**.

(g) The hypergraph $H' = (V', E', \mathcal{E}')$ is a sub-hypergraph of $H = (V, E, \mathcal{E})$ if $V' \subseteq V$, $E' \subseteq E$ and $\mathcal{E}'$ is a restriction of $\mathcal{E}$ to $E'$.

We say that sub-hypergraph $H'$ is **full** if

$$E' = \{ e \in E \mid \mathcal{E}(e) \subseteq V' \}. \quad (542)$$

In this case, we also say that $V'$ **induces** the sub-hypergraph $H'$.

(h) Unlike the trivial group $\{ e \}$ or empty ordered set, which are unique up to an isomorphism, there is no single agreed upon graph called the “trivial hypergraph”.

An unambiguous concept is that of an **edgeless hypergraph**, in which the set of hyperedges is empty, but the set of vertices may or may not be empty. Every hypergraph $H : E \to V$ has $2^{\text{ord}(V)}$ edgeless sub-hypergraphs (one for each subset of $V$).

The **order-zero hypergraph** $H : \emptyset \rightrightarrows \emptyset$ is the unique hypergraph that is a sub-hypergraph of all others.

The terms **empty hypergraph**, **null hypergraph** and **trivial hypergraph** may refer to either edgeless graphs or the order-zero graph, depending on the author and the situation.

**Proposition 1301.** Every family of nonempty sets $\mathcal{A}$ induces a hypergraph as follows:

- The set of vertices is $\bigcup \mathcal{A}$.
- The set of hyperedges is $\mathcal{A}$ itself.

**Proof.** The hypergraph is given by the function

$$H : \mathcal{A} \rightrightarrows \bigcup \mathcal{A}$$

$$\mathcal{E}(A) := A,$$

which is better represented as the binary relation

$$\{(A, x) \mid A \in \mathcal{A} \text{ and } x \in A\}. \quad \Box$$

**Example 1302.** The hypergraph $(541)$ is a counterexample to the converse of proposition 1301. It cannot be represented as a family of sets because the edges $e_2$ and $\hat{e}_2$ are parallel and would be represented by the same set.

Another counterexample is a hypergraph with a nonempty vertex set and no hyperedges. The union of the endpoints of all hyperedges is empty and is thus only a strict subset of the vertices.
Definition 1303. Suppose that we are given a Grothendieck universe $\mathcal{U}$, which is safe to assume to be the smallest suitable one as explained in definition 1112.

We denote the category of $\mathcal{U}$-small hypergraphs by $\mathcal{U}\text{-HypGph}$ or, if the universe is clear from the context, simply by $\text{HypGph}$. See definition 1118 for a further discussion of universes and categories.

- The set of objects $\text{obj}(\text{HypGph})$ is the set of all $\mathcal{U}$-small hypergraphs, i.e. the hypergraphs $H = (V, E, \mathcal{E})$ such that $V$ and $E$ are both members of $\mathcal{U}$.

- The set of morphisms $\text{HypGph}(H, G)$ from $H$ to $G$ is the set hypergraph homomorphisms. Given two hypergraphs $H : E_H \rightarrow V_H$ and $G : E_G \rightarrow V_G$, a hypergraph homomorphism is a pair of functions

$$
\begin{align*}
  f_V : V_H &\rightarrow V_G \\
  f_E : E_H &\rightarrow E_G
\end{align*}
$$

such that, for every hyperedge $e \in E_H$, if $v \in V_H$ is an endpoint of $e$, then $f_V(v)$ is an endpoint of $f_E(e)$.

Note that “graph embedding” commonly refers to an embedding of its geometric realization, hence we will avoid the term when referring to injective hypergraph homomorphisms. Furthermore, it should be clarified whether we mean “injective on vertices” or “injective on edges”, which is an important distinction in category theory — see definition 1150.

- The composition of the morphisms

$$(f_V, f_E) \in \text{HypGph}(A, B), \quad (g_V, g_E) \in \text{HypGph}(B, C)$$

is the morphism

$$(g_V \circ f_V, g_E \circ f_E) \in \text{HypGph}(A, C).$$

- The identity morphism on the hypergraph $H = (V, E, \mathcal{E})$ is the pair of identity functions $(\text{id}_V, \text{id}_H)$.

Definition 1304. Let $H = (V, E, \mathcal{E})$ be a finite hypergraph. Its incidence matrix

$$M = \{M_{ve}\}_{v \in V, e \in E}$$

has elements

$$M_{ve} := \begin{cases} 
  1, & v \text{ is incident to } e \\
  0, & \text{otherwise.}
\end{cases}$$

Compare this definition to definition 1325.
Example 1305. The incidence matrix of the hypergraph (541) is

\[
\begin{pmatrix}
 1 & 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix}
\]

Every column corresponds to a hyperedge and its nonzero elements are the endpoints of the hyperedge.

Definition 1306. Let \( H = (V, E, \mathcal{E}) \) be a hypergraph. We introduce several vector spaces that allow us to study hypergraphs using linear algebra.

(a) The vertex space \( F_2^V \) is the function space \( F_2^V \), where \( F_2 \) is the two-element field.

Every subset \( U \subseteq V \) of vertices induces a unique vector \( \overrightarrow{U} = \{\overrightarrow{u} \mid u \in U\} \) in the vertex space \( F_2^V \) such that

\[
\overrightarrow{u} := \begin{cases} 
1, & u \in A \\
0, & u \notin A 
\end{cases}
\]

This vector is called the characteristic vector of \( U \). Conversely, every vector in \( F_2^V \) induces a set of vertices. If \( U \) consists of a single vertex \( u \), we write \( \overrightarrow{u} \) rather than \( \{u\} \).

Without a choice of ordering of \( V \), there is no canonical basis in \( F_2^V \). Thus, even for finite graphs, we cannot in general regard the vectors of \( F_2^V \) as ordered tuple.

(b) Analogously, the hyperedge space \( F_2^E \) is the function space \( F_2^E \). Every subset of \( E \) induces a unique characteristic vector in \( F_2^E \) and vice versa.

The space is motivated by undirected paths and their characteristic vectors.

Proposition 1307. Let \( V = \{1, \ldots, m\} \) and \( E = \{1, \ldots, n\} \). Then there is a bijection between the \( m \times n \) matrices over \( F_2 \) and the hypergraphs \( H = (V, E, \mathcal{E}) \).

Proof. The incidence matrix of every hypergraph is a matrix in \( F_2^{m \times n} \).

Conversely, let \( M = \{M_{ve}\} \in F_2^{m \times n} \) and define the hypergraph \( \hat{H} := (V, E, \mathcal{E}) \), where

\[
\mathcal{E} : E \rightarrow V \\
\mathcal{E}(e) := \{v \in 1, \ldots, n \mid M_{ve} = 1\}.
\]

Then the adjacency matrix of \( H \) is \( M \).

Definition 1308. Let \( H = (V, E, \mathcal{E}) \) be a hypergraph. We say that a set \( T \subseteq V \) of vertices is a transversal of \( H \) if it is incident to every edge of \( H \).
**Example 1309.** Every hypergraph has at least one transversal since the vertex set itself a transversal.

**Definition 1310.** A transversal $T$ of the hypergraph $H = (V, E, \mathcal{E})$ is said to be **minimal** if any of the following equivalent conditions hold:

(a) $T$ is a minimal element under set inclusion in the set of all transversals of $H$.

(b) For every vertex $v$ in $T$ there exists a hyperedge $e_v \in E$ such that $T \cap \mathcal{E}(e_v) = \{v\}$.

**Proof.**

**Proof that 1310 (a) implies 1310 (b).** Let $T$ be minimal under inclusion among all transversals.

Fix a vertex $v \in T$. Since $T$ is minimal, the set $T \setminus \{v\}$ is not a transversal. So there exists a hyperedge $e_v \in E$ such that $(T \setminus \{v\}) \cap \mathcal{E}(e_v) = \emptyset$.

Now since $T$ is a transversal for $H$, the set $T \cap \mathcal{E}(e_v)$ is nonempty and thus

$$T \cap \mathcal{E}(e_v) = \left( (T \setminus \{v\}) \cup \{v\} \right) \cap \mathcal{E}(e_v) \overset{(505)}{=} \left( (T \setminus \{v\}) \cap \mathcal{E}(e_v) \right) \cup \{v \cap \mathcal{E}(e_v)\} = \emptyset$$

$$= \{v\}.$$

**Proof that 1310 (b) implies 1310 (a).** Now suppose that for every vertex $v \in T$ there exists a hyperedge $e_v \in E$ such that $T \cap \mathcal{E}(e_v) = \{v\}$.

Suppose that $T$ is not minimal. Then there exists some vertex $w \in T$ be such that $T \setminus \{w\}$ is a transversal. But our assertion gives us an edge $e_u \in E$ such that $T \cap \mathcal{E}(e_u) = \{w\}$. Clearly the set $T \setminus \{w\}$ cannot be a transversal of $H$ since

$$(T \setminus \{w\}) \cap \mathcal{E}(e_u) = \emptyset.$$

This contradiction proves that $T$ is minimal under set inclusion. \qed

**Example 1311.** We will give an example of a hypergraph without a minimal transversal.

For every nonnegative integer $n$ define the set

$$e_n := \{n, n + 1, n + 2, \ldots\}.$$

Let $H$ be the hypergraph whose hyperedges are $e_1, e_2, \ldots$. The set of vertices is $e_1$, hence $e_1$ is also a transversal.

Now assume that $T$ is a minimal transversal for $H$. Since, by proposition 11, the natural numbers are well-ordered, $T$ has a minimum. Let $n_0 := \min T$.

But $T \setminus \{n_0\}$ is also a transversal because each hyperedge $e_n$ intersects $T$ at infinitely many points besides $n_0$.

The obtained contradiction shows that $H$ has no minimal transversal.
**Theorem 1312** (Hypergraph minimal transversal existence). Every hypergraph has a minimal transversal.

Within ZF, this theorem is equivalent to the axiom of choice — see theorem 990 (f).

*Proof.* The proof is merely a translation of the axiom of choice into the language of hypergraphs.

**Proof that the axiom of choice implies minimal transversal existence.** Let $H = (V, E, \mathcal{E})$ be a hypergraph. Then

$$\{\mathcal{E}(e) \mid e \in E\}$$

is a family of nonempty sets and thus there exists a set $B \subseteq V$ such that $B \cap \mathcal{E}(e)$ for every hyperedge $e \in E$. This is a minimal transversal by definition 1310 (b).

**Proof that minimal transversal existence implies axiom of choice.** Let $\mathcal{A}$ be an arbitrary family of nonempty sets. Let $H$ be the hypergraph induced by $\mathcal{A}$. Let $T$ be a minimal transversal of $H$. Then, by definition, for every hyperedge $e \in \mathcal{A}$, the intersection $T \cap \mathcal{E}(e)$ is a singleton set. Hence, $T$ is the image of a choice function for $\mathcal{A}$. □
16.4. Undirected graphs

Remark 1313. Unfortunately, the term “graph” has at least several distinct established meanings:

- The graph of a valued function (or relation).
- An undirected graph, with or without loops or parallel edges.
- A directed graph, which may be a simple directed graph or a quiver.

The graph of a function is fundamentally different from the others, which we call “combinatorial graphs”. Although in definition 1314 we define undirected graphs as functions, the concepts are distinct and the formalisms are related by accident.

Definition 1314. An undirected multigraph is a hypergraph $G = (V, E, \mathcal{E})$ in which all hyperedges are either edges or loops, i.e. every hyperedge $e$ has either one or two endpoints. We will use the term “edge” to include loops, although this in general varies with the author.

Figure 37 illustrates the concepts from this definition. The figure is not merely illustrative — it is a graph embedding.

![Graph Embedding](545)

Figure 37: An undirected multigraph, which becomes simple after removing the dashed edges.

(a) A common assumption is that an undirected multigraph has no loops and no parallel edges. In this case, an established terminology is simple undirected graph.

For a simple undirected graph, we often write $G = (V, E)$, where $E$ is a set of unordered pairs $\{u, v\}$ of distinct vertices. We can also regard simple undirected graphs as symmetric directed graphs — see remark 1328.

(b) The subgraphs of $G$ are the precisely sub-hypergraphs of $G$. In particular, the full subgraphs are the full sub-hypergraphs, that are maximal among the subgraphs with the same vertices.

(c) For a Grothendieck universe $\mathcal{U}$, we denote the category of undirected (multi)graphs by $\mathcal{U}$-$\text{MultGph}$ and $\mathcal{U}$-$\text{SimpGph}$, respectively. Both are strict subcategories of the category $\text{HypGph}$ of hypergraphs.
Example 1315. For every set $V$, the complete graph $K_V$ is the simple undirected graph with an edge between any pair of distinct vertices.

Proposition 1018 justifies using the notation $K_n$ for complete graphs of finite order.

Definition 1316. Let $G = (V, E, \mathcal{E})$ be a simple undirected graph. Its adjacency matrix

$$M = \{a_{uv}\}_{u,v \in V}$$

has elements

$$M_{uv} := \begin{cases} 1, & u \text{ is adjacent to } v \\ 0, & \text{otherwise}. \end{cases}$$

Example 1317. The adjacency matrix of the complete graph $K_4$ (fig. 38) is

$$a \ b \ c \ d$$

$$\begin{pmatrix} a & 0 & 1 & 1 \\ b & 1 & 0 & 1 \\ c & 1 & 1 & 0 \\ d & 1 & 1 & 0 \end{pmatrix}$$

Definition 1318. Let $G = (V, E, \mathcal{E})$ be an undirected multigraph. We will adapt the definition of a directed path from definition 1329. Different authors use similar terms like walk and trail and put different restrictions on them. We will not put as few restrictions as possible.
Analogously to quiver path, consider a sequence \( p = (v, e_1, e_2, \ldots) \), where \( e_1, e_2, \ldots \) are edges and \( v \) is a vertex incident to \( e_1 \).

Unlike arcs in quivers, however, edges have no head and tail. Fortunately, we can define them via mutual recursion, with the caveat that every edge may have a different head and tail with respect to another path:

- If \( e_k \) is a loop at \( u \), then \( u \) is both the head and the tail of \( e_k \).
- If \( e_1 \) is not a loop, its head is \( v \) and the tail is the other endpoint of \( e_1 \).
- If \( e_k \neq e_1 \) is not a loop, if \( e_{k-1} \) has a tail \( t(e_{k-1}) \) and if \( t(e_{k-1}) \) is an endpoint of \( e_k \), we define the head of \( e_k \) to be \( t(e_{k-1}) \) and the tail of \( e_k \) is the other endpoint of \( e_k \).

If the head and tail is defined for every edge in \( p \), and if \( e_k \neq e_{k-1} \) but \( h(e_k) = t(e_{k-1}) \) for every index \( k \), we say that \( p \) is a path.

Note that it is not sufficient for \( e_k \) and \( e_{k-1} \) to be adjacent. Consider the following scenario, in which \((v, e_1, e_2, e_3)\) is not a path because the tail of \( e_2 \) is not incident to \( e_3 \):

\[
\begin{array}{c}
\bullet \quad e_2 \\
\bullet \\
\bullet \\
v
\end{array}
\]

We introduce the following terminology:

(a) The **head** of the entire path \( p = (v, e_1, \ldots, e_n) \) is \( h(p) := v \) and the **tail** is \( t(p) := t(e_n) \). Note that the tail is only defined for finite paths.

(b) The **empty path** at \( v \) is the path \( p = (v) \). We call it empty because it has no edges. It is not an empty sequence because we want it to have a head and tail. The primary motivation for different empty paths are the identity morphisms in the free category for a quiver defined in definition 1337.

(c) The **length** \( \text{len}(p) \) of a path \( p \) is simply the number of edges. Hence, empty paths have length zero.

(d) The **domain** \( \text{dom}(p) \) of \( p \) is the set of all vertices that belong to at least one edge in \( p \). We say that the path visits each member of \( \text{dom}(p) \).

The domain of the path (548) is the set \( \{e_2, e_4, e_6\} \).

Note that

\[
\text{card}(\text{dom}(p)) + 1 \leq \text{len}(p) \quad (550)
\]

in general since a vertex can be an endpoint of many edges.
Similarly to concatenation of words defined in definition 783 (f), the concatenation of the paths

\[ p = (v, e_1, \ldots, e_n) \]
and
\[ q = (t(p), e_{n+1}, e_{n+1}, \ldots) \]
is simply the path
\[ p \cdot q := (v, e_1, e_2, \ldots, e_n, e_{n+1}, e_{n+1}, \ldots). \]

Note that the head of \( q \) must be the tail of \( p \) and that \( p \) is necessarily finite.

We say that the subsequence \( q \) of \( p \) is a subpath of \( p \) if the edges of \( q \) are consecutive in \( p \), i.e. there exists some index \( n \geq 0 \) such that \( q_k = e_{n+k} \) for \( 0 < k < \text{len}(q) \).

In (548), the subsequence \((a, e_2, e_4)\) is a subpath and \((a, e_2, e_6)\) is not.

A nonempty path \( p \) is a cycle if \( h(p) = t(p) \). Cycles are also called closed paths.

Consider the graph (548). The path \((a, e_2, e_4, e_1)\) is a cycle and so is \((a, e_2, e_2)\). We have forbidden for the same edge to appear twice consecutively, which means that \((a, e_2, e_2)\) is not a cycle because it is not a valid path.

A graph without cycles is called acyclic.

The path \( p \) is simple if no subpath is a cycle.

An alternative characterization is that every vertex in \( \text{dom}(p) \) is visited only once, i.e. for every vertex in \( \text{dom}(p) \), there exist at most two edges of the path which are incident to it.

Another characterization is that equality holds in (550).

If the chain (548) is finite of length \( n \), we define its converse as
\[ p^{-1} := (t(p), e_n, e_{n-1}, \ldots, e_2, e_1). \]

The characteristic vector \( \vec{p} \) of a path \( p \) is defined as the characteristic vector of \( \text{dom}(p) \) in the edge space \( \mathbb{F}_2^E \).

Definition 1319. Let \( G = (V, E, \mathcal{E}) \) be an undirected multigraph.

(a) We say that the vertices \( u \) and \( v \) are connected if there exists a path from \( u \) to \( v \).
In relation to definition 1331, we also say that \( v \) is reachable from \( u \) and vice versa.

(b) Connectedness is obviously an equivalence relation in \( E \). Denote it by \( \sim \). For each vertex \( v \), the equivalence class \([v]\) of all vertices reachable from \( v \) is called a connected component.

The connectivity number of \( G \) is the cardinality \( \text{card}(V/\sim) \) of the quotient set.
If \( G \) has only one connected component, we say that it itself is connected. Equivalently, \( G \) is connected if every vertex is reachable from every other vertex.
(c) The **condensation** of \( G \) is the edgeless graph with vertices \( V/\sim \).

This concept is useless for undirected graphs but important for **quivers** — see **definition 1331**.

Compare this definition to **definition 1334**.
16.5. Quivers

**Definition 1320.** A **quiver** or **directed multigraph** is an extension of **undirected multigraphs.** A quiver consists of the following:

(a) A set \( V \) of **vertices**.

(b) A disjoint from \( V \) set \( A \) of **arcs.** It is conventional to call them arcs or even directed edges rather than simply edges.

(c) A **head** function \( h : A \rightarrow V \).

(d) A **tail** function \( t : A \rightarrow V \).

(e) We say that the arc \( a \) is a **successor** of \( b \) and that \( b \) is a **predecessor** of \( a \) if \( t(a) = h(b) \). We also say that \( a \) and \( b \) are **consecutive.**

   For vertices, we say that \( w \) is a successor of \( v \) and that \( w \) is consecutive to \( w \) if there exists an arc from \( u \) to \( v \).

(f) For each quiver \( Q = (V, A, h, t) \), we have an “underlying” undirected multigraph

\[
G : A \rightarrow V, \\
G(a) := \{h(a), t(a)\}. 
\]

(551) All properties of \( G \) are inherited by \( Q \), although some of them do not really fit or need adaptation.

We denote this graph by \( U(Q) := G \) and will regard it as a forgetful functor — see **definition 1323.**

(g) One adaptation is that the arcs \( a \) and \( b \) are called **opposing** they are successors of each other and the term “parallel” is reserved for arcs which are parallel in the sense of **definition 1300 (b)** but not opposing.

(h) If there are no parallel and opposing arcs or loops, we call the quiver **simple.** The term **simple directed graph** is more commonly used.

   For a simple directed graph, we often write \( G = (V, E) \), where \( E \) is an **irreflexive** binary relation.

   We say that the simple directed graph is **symmetric** if this relation is symmetric. For many purposes, symmetric directed graphs can be regarded as undirected — see **remark 1328.**

(i) The **subquivers** of \( Q \) are the precisely **subgraphs** of \( G \). In particular, the **full subquivers** are the **full subgraphs**, that are maximal among the subgraphs with the same vertices.
Example 1321. A very simple example of an infinite quiver is the transitive reduction of the positive integers, i.e. the simple directed graph

$$1 \rightarrow 2 \rightarrow 3 \rightarrow \cdots \rightarrow n - 1 \rightarrow n \rightarrow n + 1 \rightarrow \cdots$$

Since the graph is simple, we have

$$\text{deg}(n) = \begin{cases} 1 = 0 + 1, & n = 0, \\ 2 = 1 + 1, & n > 0 \end{cases}$$

and thus \(\text{deg}(G) = 2\).

The categorical diagrams corresponding to this quiver is used to define direct limits in definition 1211 (a).

Another related graph is based on the negative integers:

$$\cdots \rightarrow -n \rightarrow -n + 1 \rightarrow \cdots \rightarrow -3 \rightarrow -2 \rightarrow -1$$

The categorical diagrams corresponding to this quiver is used to define inverse limits in definition 1211 (b).

Finally, the union of the two with zero added gives us the following quiver:

$$\cdots \rightarrow -n \rightarrow \cdots \rightarrow -1 \rightarrow 0 \rightarrow 1 \rightarrow \cdots \rightarrow n \rightarrow \cdots$$

All three graphs are infinite but locally finite.

Definition 1322. Suppose that we are given a Grothendieck universe \(\mathcal{U}\), which is safe to assume to be the smallest suitable one as explained in definition 1112.

We denote the category of \(\mathcal{U}\)-small quivers by \(\mathcal{U}\text{-Quiv}\) or, if the universe is clear from the context, simply by \(\text{Quiv}\). See definition 1118 for a further discussion of universes and categories.

A more “categorical” definition, in which this category arises naturally, is given in example 1155.

For simple directed graphs, we denote the category using \(\text{SimpQuiv}\).
• The set of objects \( \text{obj}(\text{Quiv}) \) is the set of all \( \mathcal{U} \)-small quivers, i.e. the quivers \( Q = (V, A, h, t) \) such that \( V \) and \( A \) are both members of \( \mathcal{U} \).

• The set of morphisms \( \text{Quiv}(Q, R) \) from \( Q \) to \( R \) is the set quiver homomorphisms. Given two quivers \( Q = (V_Q, A_Q, h_Q, t_Q) \) and \( R = (V_R, A_R, h_R, t_R) \), a quiver homomorphism is a pair of functions

\[
\begin{align*}
  f_V : V_Q &\to V_R \\
f_A : A_Q &\to A_R
\end{align*}
\]

such that

\[
\begin{align*}
  h_R \circ f_A &= f_V \circ h_Q \\
t_R \circ f_A &= f_V \circ t_Q
\end{align*}
\]

See example 1155 for a categorical justification of this definition, including reducing the above conditions to the diagram (385).

Clearly every quiver homomorphism is a hypergraph homomorphism.

Note that “quiver embedding” commonly refers to an embedding of its geometric realization, hence we will avoid the term when referring to injective quiver homomorphisms. Furthermore, it should be clarified whether we mean “injective on vertices” or “injective on arcs”, which is an important distinction in category theory — see definition 1150.

• The composition of quiver morphisms is pointwise composition — the same as composition of hypergraph homomorphisms.

• The identity morphism on the quiver \( Q = (V, A, h, t) \) is the pair of identity functions \((\text{id}_V, \text{id}_A)\).

[\text{nLa17}] \textbf{Definition 1323.} Given an undirected (multi)graph \( G = (V, E, \mathcal{E}) \), constructing a quiver requires a choice of head and tail for every edge. We can view undirected graphs as equivalence classes of quivers. A head function is merely a choice function on the set \( \{\mathcal{E}(e) \mid e \in E\} \), and a choice function can be provided by the axiom of choice even for infinite graphs.

For a given choice function \( c \), we denote by \( O_c(G) = (V, E, h, t) \) the quiver obtained by identifying the head \( h(e) \) with \( c(\mathcal{E}(e)) \) and the tail \( t(e) \) with the other endpoint of \( e \), if one exists, and with \( h(e) \) for loops. We call \( O_c(G) \) the \textbf{orientation} induced by the choice function \( c \).

Note that \( O_c \) is not left adjoint to \( U \) — see example 1192 (e).

\textbf{Remark 1324.} As we shall see, the adjacency and incidence of a quiver can be studied using linear algebra via adjacency and incidence matrices, while the connectedness can be studied using topology via graph embeddings.

[\text{GM84}] \textbf{Definition 1325.} Let \( Q = (A, E, h, t) \) be a finite quiver. Its \textbf{incidence matrix}

\[
M = \{M_{va}\}_{v \in V, a \in A}
\]

ch. 1, sec. 2.1

691
has elements

\[ M_{\text{in}} := \begin{cases} 
1, & v = h(e) \\
-1, & v = h(e) \\
0, & \text{otherwise.}
\end{cases} \]

Compare this definition to definition 1304.

**Example 1326.** The incidence matrix of the simple directed graph (552) is

\[
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\]

It can be read column-by-column. Every column contains exactly two nonzero elements whose rows correspond to the head (positive) and tail (negative).

To obtain the incidence matrix for the underlying undirected graph (545), we need to simply flip the sign of the boxed elements above.

The adjacency matrix is,

\[
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\]

where the boxed elements are nonzero only in the adjacency matrix for (545) and not (552).

The matrix can be read either column-by-column or row-by-row.

- The \(v\)-th column lists the vertices \(u\) such that there is an arc from \(u\) to \(v\).
- The \(u\)-th row lists the vertices \(v\) such that there is an arc from \(v\) to \(u\).

**Proposition 1327.** Every square finite matrix over the two-element field \(\mathbb{F}_2\) corresponds to a quiver without parallel arcs.

If the matrix is symmetric, we obtain its underlying undirected graph.

**Proof.** Trivial.
Remark 1328. It is sometimes convenient to conflate symmetric simple directed graphs with simple undirected graphs.
This is partially justified by proposition 1327.

Definition 1329. Let $Q = (V, A, h, t)$ be a quiver.

(a) A directed path or simply path in $Q$ a finite or infinite sequence of consecutive arcs with a specified head vertex. Formally, it is a sequence $p = (v, e_1, e_2, ...)$, where $h(e_k) = t(e_{k-1})$ for every index $k$.
This is a simplification of definition 1318 since every arc has a head and tail. It is also too restrictive for some purposes.
The definitions of path endpoints, empty path, length, domain, concatenation and characteristic vector of a directed path are inherited from definition 1318.

(b) A directed cycle is a directed path whose endpoints coincide. The term circuit is also used, for example in [GM84, ch. 1, sec. 3.2]. A quiver without cycles is called acyclic and a simple acyclic directed graph is commonly abbreviated as DAG.

(c) Similarly to definition 1318 (h), we say that a directed path is simple if it contains no directed cycles.

(d) If $p = (v, e_1, e_2, ...) is a path in the undirected multigraph $U(Q)$,

$$D(p) := (v, e_1, e_2, ...)$$

is a sequence of adjacent arcs in $Q$ (with a specified head vertex).
The head and tail of $D(p)$ are defined as the head and tail of $p$. This definition obviously differs from definition 1329.
We say that the arc $e_k$ of $D(p)$ is positively oriented if the head $h_k$ of the edge $e_k$ in $Q$ equals the head $h(e_k)$ of the arc $e_k$ in $D(p)$. If $e_k$ is not positively oriented, we say that it is negatively oriented.
We conflate $p$ and $D(p)$ where this does not cause confusion.
Undirected paths are also called chains in [GM84, ch. 1, sec. 3.2].

Example 1330. An example of an infinite directed path is the reduced positive integer graph (553) regarded as a path of the reduced integer graph (555). It is also simple since the degree of (553) is 2.
Now consider again the quiver (552). The solid lines in (561) describe a path.
• This path corresponds to the sequence
  \[ p = (a, a \to c, c \to d, d \to f). \]

  ![Diagram](image)

• Its characteristic vector is
  \[ \vec{p} = (0, 1, 0, 1, 0, 1, 0) \.
  \]

Furthermore, \( p \) can be identified from the characteristic vector.

• The endpoints are \( h(p) = h(e_1) = a \) and \( t(p) = t(e_7) = f \).

• It is a simple path because \( \deg(a) = \deg(f) = 1 \) and \( \deg(c) = \deg(d) = 2 \) in the induced subgraph.

• The converse undirected path
  \[ p^{-1} = (f, d \to f, c \to d, a \to c), \]
  is not a directed path because every arc is negatively oriented.

**Definition 1331.** Let \( Q = (V, A, h, t) \) be a quiver. We say that the vertex \( w \) is **reachable** from \( v \) if there exists a directed path from \( v \) to \( w \).

Define the equivalence relation

\[ v \sim w \iff v \text{ is reachable from } w \text{ and } w \text{ is reachable from } u. \]

Now define the binary relation \( \tilde{A} \) on the quotient set \( \tilde{V} := V/\sim \) as

\[ ([u], [v]) \in \tilde{A} \iff v \text{ is reachable from } w \text{ but not vice versa}. \]

It is well-defined because if \( ([v_1], [w_1]) \in \tilde{A}, v_2 \in [v_1] \) and \( w_2 \in [w_1] \), then by the transitivity of reachability we have that \( w_2 \) is reachable from \( v_2 \) and thus \( ([v_2], [w_2]) \in \tilde{A} \).

As discussed in **definition 1320 (h)**, irreflexive relations over sets can be regarded as simple directed graphs. Therefore, the pair \( \tilde{G} := (\tilde{V}, \tilde{A}) \) is a directed graph. It is called the **condensation** of the quiver \( Q \).

Compare this definition to **definition 1319 (c)**.

**Example 1332.** The condensation of the quiver in (552) is (isomorphic to) the graph itself. If we add the arc \( f \to a \) to (552), the condensation would be an edgeless quiver with a single vertex.
We can add new arcs $e_8$ and $e_9$ to the quiver in (552) to make the example more interesting:

![Quiver Diagram](image)

Its condensation is

$$\{a\} \xrightarrow{\{e_1,e_2\}} \{b,c,d,e\} \xrightarrow{\{e_6,e_9\}} \{f\}$$

These are the **strongly connected components** of (562).

**Proposition 1333.** The condensation of a quiver is a directed acyclic graph.

**Proof.** Let $\tilde{G} = (\tilde{V}, \tilde{E})$ be the condensation of the quiver $Q = (V,A,h,t)$. We have already discussed in definition 1331 that it is a directed graph. It remains to show that it is acyclic.

Aiming at a contradiction, suppose that there exist cosets $[u]$ and $[v]$ and a path

$$([u] \rightarrow [w_1], \ldots, [w_n] \rightarrow [v])$$

in $\tilde{G}$ that connects them. We can easily prove by induction that $v$ is reachable from $u$, thus contradicting the definition of $\tilde{A}$.

Therefore, $\tilde{G}$ is acyclic.

**Definition 1334.** Let $Q = (V,E,h,t)$ be a quiver and let $\tilde{G} = (\tilde{V}, \tilde{E})$ be its condensed directed graph.

(a) For every coset $[v]$ in $\tilde{V}$, the subquiver of $Q$ induced by the vertices $[v]$ is called a **strongly connected component** of $Q$.

The **strong connectivity number** of $Q$ is the cardinality $\text{card}(\tilde{V})$.

If $Q$ has only one strongly connected component, we say that it itself is **strongly connected**.

(b) Similarly, the subquiver $Q'$ of $Q$ is called a **weakly connected component** if it is a connected component, in the sense of definition 1319 (b), of the underlying undirected multigraph $U(Q)$.

The **weak connectivity number** of $Q$ is the connectivity number of $U(Q)$.

If $Q$ has only one weakly connected component, we say that it itself is **weakly connected**.

This definition generalizes symmetric and transitive closures of binary relations not commuting — see example 963.

Compare this definition to definition 1319.
Proposition 1335. Let $Q$ be a quiver.

(a) $Q$ is strongly connected if and only if there exists a directed path connecting every pair of vertices.

(b) $Q$ is weakly connected if and only if there exists an undirected path connecting every pair of vertices.

Proof. Trivial. □

Remark 1336. We can regard any set $A$ in the sense of ZFC as the simple directed graph $(A, \in)$.

The axiom of foundation (via proposition 993) implies that the relation $\in$ is well-founded.

In the terminology if graph theory, this well-foundedness means that, for every vertex $v$, there exists no infinite path that ends with $v$.

This implies that the graph is acyclic. For finite graphs the converse holds, but an infinite graph this is not so. For example, the chain (553) is acyclic but not well-founded.

Well-founded graphs are important for theorem 997 (Well-founded induction).

Definition 1337. Let $Q = (V, h, t)$ be a quiver. We define the free category $F(Q)$ generated by $Q$ as follows:

- The set of objects $\text{obj}(F(Q))$ is the set of vertices $V$.

- The set of morphisms $F(Q)(v, w)$ is the set of all paths from $v$ to $w$.

- The composition of the morphisms $p$ and $q$ with $h(q) = t(p)$ is the concatenation

  \[ q \circ p = p \cdot q. \]

- The identity morphism on the vertex $v$ is the empty path at $v$. This is the primary motivation for having a distinct empty path at every vertex.

Since $F(Q)$ is already defined for every quiver $Q$, if we also define how it acts on quiver homomorphisms, this will make $F$ a functor from the category $\mathcal{U}$-$\textbf{Quiv}$ of $\mathcal{U}$-small quivers to the category $\mathcal{U}$-$\textbf{Cat}$ of $\mathcal{U}$-small categories, for every Grothendieck universe $\mathcal{U}$ containing $Q$.

For every $\mathcal{U}$-small category $\mathcal{C}$ and every quiver homomorphism $(g_V, g_A) : Q \to U(\mathcal{C})$, we define the following functor:

\[ G : F(Q) \to \mathcal{C}, \]
\[ G(v) := g_V(v) \]
\[ G(v, a_1, \ldots, a_n) := \begin{cases} \text{id}_v, & n = 0, \\ g_A(a_n) \circ G(v, a_1, \ldots, a_{n-1}), & n > 0. \end{cases} \quad (564) \]

The functor $G$ “evaluates” paths in $Q$ inside $\mathcal{C}$. Put $F(f_V, f_A) := G$. Parameterized on $(f_V, f_A)$, this defines $F$ is a functor from $\mathcal{U}$-$\textbf{Quiv}$ to $\mathcal{U}$-$\textbf{Cat}$.

We will see in example 1192 (d) that $F$ is actually left adjoint to the forgetful functor $U$.

The new identity loops are an important part of this adjunction.

696
16.6. Trees

In this section, we will regard the edges of simple undirected graphs as sets of unordered tuples and edges of simple directed graphs as sets of ordered tuples.

**Definition 1338.** The possibly infinite simple undirected graph \( T = (E, V, \mathcal{E}) \) is called a tree if any of the following equivalent conditions hold:

(a) \( T \) is connected and acyclic.

(b) \( T \) is maximally acyclic, meaning that adding an edge between existing vertices would create a cycle.

(c) \( T \) is minimally connected, meaning that removing an edge will make the graph disconnected.

(d) For every pair of vertices in \( T \), there exists a unique path connecting them.

This condition motivates the definition of arborescences.

It is conventional to use the term node for the vertices of a tree.

**Proof.**

**Proof that 1338 (a) implies 1338 (b).** Suppose that \( T \) is connected and acyclic. There is always an edge not in \( T \) because the complete graph \( K_V \) contains cycles and \( T \) does not. Let \( \{u, v\} \) be an edge not in \( T \) and let \( T' \) be the supergraph that adjoins only this new edge.

Since \( T \) is connected, there exists a path \( p \) connecting \( u \) and \( v \). Appending the edge \( \{u, v\} \) to this path creates a cycle in \( T' \).

Therefore, \( T' \) is not acyclic.

**Proof that 1338 (b) implies 1338 (a).** Suppose that \( T \) is maximally acyclic. We will show that it is connected.

Suppose that \( T \) is not connected. Then there exist vertices \( u \) and \( v \) with no path between them. Let \( T' \) be the supergraph that adjoins only the edge \( \{u, v\} \).

Since there is no path between \( u \) and \( v \) in \( T \), \( T' \) is also acyclic. But this contradicts the maximality of \( T \).

Therefore, \( T \) is connected.

**Proof that 1338 (a) implies 1338 (c).** Suppose that \( T \) is connected and acyclic. Let \( \{u, v\} \) be any edge of \( T \) and let \( T' \) be the subgraph that does not contain this edge.

Suppose that \( T' \) is connected. Let \( w \) be any vertex and let \( p \) be a path in \( T' \) from \( u \) to \( w \). Then adding \( \{u, v\} \) creates a cycle in \( T \), which contradicts our assumption that \( T \) is acyclic.

Therefore, \( T' \) is not connected.

**Proof that 1338 (c) implies 1338 (a).** Suppose that \( T \) is minimally connected. We will show that it is acyclic.

Suppose that \( T \) has a cycle \( p \). We can thus remove any edge of \( p \) from \( T \) and the resulting subgraph will also be connected. This contradicts the minimality of \( T \).

Therefore, \( T \) is acyclic.
Proof that 1338 (a) implies 1338 (d). Suppose that $T$ is connected and acyclic. Since $T$ is connected, there exists at least one path between any two vertices. Since it is acyclic, this path must be unique because otherwise we could easily create a cycle by joining two such paths.

Proof that 1338 (d) implies 1338 (a). Suppose that, for every pair of vertices in $T$, there exists a unique path connecting them. It is clear that $T$ is connected.

Suppose that $T$ has a cycle $p$ from $u$ to $u$ passing through $v$. We thus obtain two paths from $u$ to $v$ — one not containing the first edge in $p$ and one not containing the last edge. But this contradicts our assumption that any two paths from $u$ to $v$ are equal.

Therefore, $T$ is acyclic. □

Example 1339. The reduced infinite integer graphs (553), (554) and (555) from example 1321 are simple examples of infinite trees.

Examples of finite trees are proof trees from section 12.7 (Deductive systems) and concrete and abstract syntax trees from formal grammar.

Definition 1340. Let $T = (V, A)$ be a simple directed graph. Suppose that there exists a vertex $r \in V$ such that for every other vertex $v \in V$, there exists a unique directed path from $r$ to $v$. The pair $(G, r)$ is called a (directed) arborescence. It is often more convenient to consider the triple $T = (V, A, r)$ instead. The vertex $r$ is called the root of the arborescence.

(a) A tree with a distinguished vertex $r$ is called a rooted tree.

For every tree $T$, as a consequence of definition 1338 (d), every vertex $r$ of $T$ induces an orientation $O(T)$ of $T$ such that $O(T)$ is an arborescence.

For this reason, we identify rooted trees with their induced arborescences.

(b) The depth of a node $v$ is the length of the path from $r$ to $v$. The term level is also sometimes used but with a slightly different meaning — the greater the depth, the lower the level.

(c) If $v$ has a strictly greater depth than $u$, we say that $u$ is an ancestor of $v$ and that $v$ is a descendant of $u$.

The ancestor of $v$ at the lowest possible level is called the parent of $v$. If $u$ is a parent of $v$, $u$ is a child of $v$. If a node has no children, we say that it is a leaf node.

Finally, if $u$ and $v$ are on the same level, we call them siblings.

(d) The height or depth of the entire tree $T$, if it exists, is the supremum of the depths of all nodes.

(e) The width or breadth of the entire tree $T$, if it exists, is the supremum of the number of siblings among all vertices. It is equal to the degree of $T$ minus 1.

We use terminology similar to definition 953 (c), e.g. “binary tree”, “ternary tree”, ....
A subgraph of an arborescence that is itself an arborescence is called a sub-arborescence. The same holds for rooted subtrees.

For every node \( v \) of \( T \), we define the induced sub-arborescence of \( T \) with root \( v \) as the subgraph induced by \( V \setminus \text{dom}(p) \), where \( p \) is the unique path from \( r \) to \( v \).

**Theorem 1341** (König’s lemma). Every locally finite arborescence of infinite order contains a simple path of infinite length.

**Proof.** Let \( T = (V, A, r) \) be a locally finite infinite arborescence. We will build an infinite simple path

\[
p = (p_1, p_2, \ldots)
\]

using natural number recursion starting at one. Let \( c \) be a choice function on the family

\[
pow(V) \setminus \{\emptyset\}.
\]

Such a choice function exists by the axiom of choice.

- Since \( T \) is locally finite, there are finitely many children of \( r \). If we suppose that the subarborescence induced by \( v \) is finite for every child \( v \) of \( r \), then we would obtain that \( T \) itself is finite, which is a contradiction. Therefore, for at least one child, the induced sub-arborescence is infinite. Using the choice function \( c \), pick one such child and denote it by \( v_1 \).

Define \( p_1 \) to be the arc \( r \to v_1 \).

- Fix \( k > 1 \). Using the same argument on the children of \( t(e_{k-1}) \), we can pick a child \( v_k \) that has an infinite induced sub-arborescence.

Define \( e_k \) to be the arc \( t(e_{k-1}) \to v_k \). The path

\[
(e_1, e_2, \ldots, e_{k-1}, e_k)
\]

is simple by construction because \( T \) is an arborescence and thus contains no directed cycles.

Thus, we have constructed an infinite simple path. \( \square \)
16.7. Graph embedding

**Definition 1342.** Let \( Q = (V, A, h, t) \) be a quiver. Our goal is to construct a **topological space** that translates the connectivity properties of \( Q \) into their topological equivalents.

Consider the **topological sum**

\[
S := \left( \coprod_{a \in A} [0, 1] \right) \sqcup \left( \coprod_{\deg(v) = 0} \{v\} \right).
\]

The space \( S \) consists of disjoint unit intervals, one for each arc, and of disjoint points, one for each isolated vertex.

We now want to glue common endpoints of arcs in \( \Sigma \). We define the function

\[
R_V : V \to \text{pow}(S),
\]

\[
R_V(v) := \begin{cases}
(v, v), & \deg(v) = 0 \\
\{(0, a) \mid h(a) = v\} \cup \{(1, a) \mid t(a) = v\}, & \deg(v) > 0
\end{cases}
\] (565)

and

\[
R_A : A \to \text{pow}(\text{pow}(S)),
\]

\[
R_A(a) := \{R_V(h(a)), R_V(t(a))\} \cup \{(x, a) \mid 0 < x < 1\}.
\] (566)

The family

\[
X := \bigcup \{R_A(a) \mid a \in A\} \cup \{R_V(v) \mid \deg(v) = 0\}.
\]

is a partition of \( S \). For each vertex, there is a single point in \( X \) (which is a set in \( S \)) and for each arc, the interior of the arc is a subset of \( X \).

We can endow the partition \( X \) with a **quotient topology** \( \mathcal{T} \). We will call the topological space \( (X, \mathcal{T}, R_V, R_A) \) endowed with the functions \( R_V \) and \( R_A \) the **geometric realization** of \( G \).

(a) For an undirected multigraph \( G = (V, E, \mathcal{E}) \), the geometric realization is any of the geometric realizations of its orientations. This construction is dependent on a choice function, but fortunately all the geometric realizations are homeomorphic as shown in proposition 1347. Hence, for a lot of purposes, we can speak of “the” geometric realization of an undirected multigraph.

(b) We will call any **continuous function** with domain \( (X, \mathcal{T}) \) a **graph drawing**. The term “graph drawing” is not standard terminology, but unfortunately non-injective continuous images of the realization have no established name.

(c) An injective graph drawing is called a **graph embedding**. Every graph can be embedded into \( \mathbb{R}^3 \) as shown in proposition 1345.

(d) If a graph can be embedded into \( \mathbb{R}^2 \), we say that it is **planar**.
(e) If a graph can be embedded into $\mathbb{R}$, we say that it is **linear**.

**Example 1343.** We will give a few examples of quiver geometric realizations.

(a) The geometric realization of an **edgeless** quiver is the discrete topological space on its vertices.

In particular, the geometric realization of the **order-zero** quiver (without any arcs and edges) is the empty topological space.

(b) Consider the reduced positive integer graph (553). We start with $\aleph_0$ copies of $[0, 1]$ and glue both ends of each of them except for the first. Thus, we obtain (a space homeomorphic to)

$$\bigcup_{k \geq 0} [k, k+1] = [0, \infty).$$

Therefore, (553) is a **linear** graph.

(c) The graph with vertices $V = \{a, b, c\}$ and arcs $\{a \rightarrow b, b \rightarrow c, c \rightarrow a\}$ is more subtle.

We start with three copies of the interval $[0, 1]$, depicted in (567) as upward-pointing arrows, and use dashed lines to connect the endpoints that we want to glue together.

After contracting the dashed lines, we obtain a topological space that can easily be **embedded** into $\mathbb{R}^2$. An obvious embedding corresponds to “pulling up” $e_2$ and $e_3$:

This is only one possible embedding of the geometric realization. It is sufficient, however, for proving that the graph is **planar**. The underlying undirected graph is the **complete graph** $K_3$, hence we have shown that $K_3$ is also planar.
(d) Figure 38 shows that the complete graph $K_4$ is planar.

This is not-at-all obvious from its geometric realization, however.

![Diagram of $K_4$](569)

This example shows that constructing embeddings can be a tedious task.

**Proposition 1344.** If finite quiver is linear, it has degree at most 2.

**Proof.** Let $Q = (V, A, t, h)$ be a quiver and let $(X, T, R_V, R_A)$ be its geometric realization. Let $f : X \to \mathbb{R}$ be an injective continuous function, i.e. a topological embedding.

Suppose that the vertex $v$ has degree larger than 2. It is sufficient to consider the case where $a, b$ and $c$ are distinct arcs incident to $v$.

We have 

$$R_V(v) \in R_A(a) \cap R_A(b) \cap R_A(c)$$

thus 

$$f(R_V(v)) \in f(R_A(a)) \cap f(R_A(b)) \cap f(R_A(c))$$

Since $R_A(a)$, $R_A(b)$ and $R_A(c)$ are connected, so are their images under $f$. If $f(R_A(a))$ has a point to the right of $f(R_V(v))$, then $f(R_A(b))$ must be left of $R_V(v)$ and there remains nowhere to place $R_A(c)$.

Therefore, $\deg(v) \leq 2$ and, since $v$ was arbitrary, $\deg(Q) \leq 2$. \qed

**Proposition 1345.** Every finite quiver can be embedded into $\mathbb{R}^3$.

**Proof.** Let $Q = (V, A, t, h)$ be a finite quiver of order $n$. By definition of cardinality, there exists a bijection from $n$ to $V$.

Place the vertices of $Q$ along the moment curve by using $\gamma(k)$ as the position for the $k$-th vertex of $V$. Then by Proposition 422, no four of these points are coplanar. Hence, if we connect their vertices using a straight line where there is an arc, no two lines would intersect.

Therefore, this is an embedding. \qed

**Proposition 1346.** Let $Q = (V, A, h, t)$ be a quiver and let $(X, T, R_V, R_A)$ be its geometric realization.

(a) A vertex $v \in V$ is isolated in $Q$, i.e. has degree zero, if and only if $R_V(v)$ is an isolated point of $X$.

(b) If $Q$ is locally finite, then the space $(X, T)$ satisfies the T1 separation axiom.

**Proof.**
Proof of 1346 (a). For every isolated vertex \( v \), the point \( R_v(v) = \{(v, v)\} \) is isolated by definition.

Now suppose that \( v \) is not an isolated vertex. Then \( R_v(v) \) is defined differently in (565). If there exists an arc \( a \) such that \( v = h(a) \), then \( (a, 0) \in R_v(v) \) and hence there exists no neighborhood of \( R_v(v) \) disjoint from \( R_A(a) \), hence \( R_v(v) \) is not a disjoint point of \( X \). The case \( v = t(a) \) is handled analogously.

Proof of 1346 (b). Let \( x \in X \).

If \( x = \{(t, a)\} \) for some arc \( a \) and \( 0 < t < 1 \), then \( \{x\} \) is closed because \([0, 1]\) satisfies T1 and hence \( \{t\} \) is closed in \([0, 1]\).

If \( x = R_v(v) = \{(v, v)\} \) for some isolated vertex \( v \), then \( R_v(v) \) is clopen by definition.

If \( x = R_v(v) \) for some vertex \( v \) of positive degree, then

\[
R_v(v) = \{(0, a) \mid h(a) = v\} \cup \{(1, a) \mid t(a) = v\}.
\]

Both \( \{0\} \) and \( \{1\} \) are closed in \([0, 1]\), hence \( \{0, 1\} \) is also closed in \([0, 1]\). Since \( Q \) is locally finite, \( R_v(v) \) is the union of finitely many closed sets and is thus itself closed.

Proposition 1347. Let \( G = (V, E, \mathcal{E}) \) be an undirected multigraph.

Let \((X_c, \mathcal{T}_c, R_V, R_A)\) and \((X_d, \mathcal{T}_d, R_V, R_A)\) be geometric realizations corresponding to the orientations \( O_c(G) \) and \( O_d(G) \) of \( G \).

Then \((X_c, \mathcal{T}_c)\) and \((X_d, \mathcal{T}_d)\) are homeomorphic.

Proof. Let \( h_c \) and \( h_d \) be the head functions from the quivers \( O_c(G) \) and \( O_d(G) \).

Define the function

\[
f : X_c \rightarrow X_d
\]

\[
f(x) := \begin{cases}
(0, e) & \text{if } h_d(e) = v \text{ and } t_d(e) = v, \text{ } x = R_v(v) \text{ and } \deg(v) > 0 \\
(1, e) & \text{if } h_d(e) = v \text{ and } t_d(e) = v, \text{ } x = \{(e, t)\} \text{ and } h_c(e) \neq h_d(e) \\
x & \text{otherwise.}
\end{cases}
\]

This function “reverses” the direction of some of the intervals in the construction of the realizations and fixes everything else in place. It is clearly bijective. It is also continuous because it satisfies definition 291 (f). Finally, it is a homeomorphism because the inverse function is defined in the same way by interchanging \( c \) and \( d \).

Proposition 1348. Let \( Q = (V, A, h, t) \) be a quiver and let \((X, \mathcal{T}, R_V, R_A)\) be its geometric realization.

(a) If there exists an undirected path \( p = (v, e_1, \ldots, e_n) \) from \( s \) to \( f \), then there exists a continuous path \( \gamma : [0, 1] \rightarrow X \) from \( R_V(s) \) to \( R_V(f) \).

(b) If \( Q \) is finite and if there exists a simple continuous path \( \gamma : [0, 1] \rightarrow X \) from \( R_V(s) \) to \( R_V(f) \), then there exists a simple undirected path from \( s \) to \( f \).
Proof. The case $s = f$ is trivial, and we assume that $s \neq f$.

Proof of 1348 (a). Suppose that $p$ is an undirected path from $s$ to $f$. We will use strong induction on the length of $p$ to show that there exists a continuous path between the points $R_V(s)$ and $R_V(f)$ of $X$.

Suppose that the statement holds for paths of length smaller than $n$ and let

$$p = (v, e_1, ..., e_{n-1}, e_n)$$

be a path of length $n$ from some vertex $s$ to $f$. The inductive hypothesis holds for the initial segment $(e_1, ..., e_{n-1})$ of $p$, hence there exists a continuous path $\gamma : [0, 1] \to X$ from $s$ to an endpoint of $e_{n-1}$.

- If both $e_{n-1}$ and $e_n$ are positively oriented, then $t(e_{n-1}) = h(e_n)$ and thus $\gamma$ is a continuous path from $R_V(s)$ to $R_V(h(e_n))$. We can then append to $\gamma$ the continuous path from $R_V(h(e_n))$ to $R_V(t(e_n)) = R_V(f)$ to obtain a path from $R_V(s)$ to $R_V(f)$.

- If $e_{n-1}$ is positively oriented but $e_n$ is not, then $\gamma$ is a continuous path from $R_V(s)$ to $R_V(h(e_{n-1}))$. We can then append to $\gamma$ the paths from $R_V(h(e_{n-1}))$ to $R_V(t(e_{n-1})) = R_V(t(e_n))$ and from $R_V(t(e_n))$ to $R_V(h(e_n)) = R_V(f)$.

- Similarly, if $e_{n-1}$ is negatively oriented but $e_n$ is not, then $\gamma$ is a continuous path from $R_V(s)$ to $R_V(t(e_{n-1}))$, and we can append to $\gamma$ the paths from $R_V(t(e_{n-1}))$ to $R_V(t(e_n)) = R_V(h(e_n))$ and from $R_V(h(e_n))$ to $R_V(f)$.

- Finally, if both $e_{n-1}$ and $e_n$ are negatively oriented, then $\gamma$ is a continuous path from $R_V(s)$ to $R_V(t(e_{n-1}))$, and we can append to $\gamma$ the paths from $R_V(t(e_{n-1}))$ to $R_V(h(e_{n-1})) = R_V(t(e_n))$ and from $R_V(t(e_n))$ to $R_V(h(e_n)) = R_V(f)$.

We have shown that there exists a continuous path from $R_V(s)$ to $R_V(f)$.

Proof of 1348 (b). Suppose that $Q$ is finite and let $\gamma : [0, 1] \to X$ be a continuous path from $R_V(s)$ to $R_V(f)$. We will show that there is an undirected path from $s$ to $f$.

$\gamma$ contains no isolated vertices. We have

$$\gamma^{-1}(X) = \bigcup_{a \in A} \gamma^{-1}(R_A(a)) \cup \bigcup_{\deg(v) = 0} \gamma^{-1}(R_V(v)).$$

(570)

For each arc $a$, the set $R_A(a)$ is closed as a homeomorphic image of the unit interval $[0, 1]$. From proposition 1346 (b) it follows that $\{R_V(v)\}$ is a closed set for every vertex $v \in V$. Therefore, $\gamma^{-1}(X)$ is a union of disjoint closed sets.

If we assume that $\gamma$ passes through an isolated vertex $v$, then $\gamma^{-1}(v)$ would be a nonempty closed set. Then $\text{img}(\gamma) = \{v\}$ because otherwise $[0, 1]$ would be the union of finitely many nonempty disjoint closed sets, which would contradict definition 311 (b) because $[0, 1]$ is connected.

But $\gamma$ passes through at least two vertices because $s \neq t$, and hence it doesn’t pass through any isolated vertex.
**γ contains the entirety of each arc it intersects.** Suppose that $R_A(a) \cap \text{img}(γ) = \emptyset$ for some arc $a$.

Let $l := \inf\{0 < t < 1 \mid γ(t) \in R_A(a)\}$ and $r := \sup\{0 < t < 1 \mid γ(t) \in R_A(a)\}$. The closed interval $[l, r]$ is compact, hence $γ([l, r])$ is a continuous image of a compact set and hence is itself compact.

Since $\text{int} R_A(a)$ is a subset of $γ([l, r])$ and since $γ([l, r])$ is closed, its closure $R_A(a)$ is also a subset of $γ([l, r])$.

Therefore, if an internal point of an arc belongs to $\text{img}(γ)$, so does the entire arc.

**γ contains an arc.** From (570) if follows that

$$γ^{-1}(X) = \bigcup_{a \in A} γ^{-1}(R_A(a)) = \overline{\bigcup_{a \in A} γ^{-1}(\text{int} R_A(a))}.$$

Hence, $\text{img}(γ)$ contains at least one internal point of some arc and thus the entire arc.

**γ induces an undirected path from $s$ to $f$.** We have that $γ(0) = R_V(s)$. By what we have already shown, $\text{img}(γ)$ contains no isolated points, hence $γ$ contains the set $R_A(a)$, where $a$ is some arc incident to $s$.

Suppose that $h(a) = s$. Then $p = (s, a)$ is an undirected path from $s$ to $t(a)$. If $t(a) = f$, the proof is finished. Otherwise, define

$$x_0 := \sup\{x \in [0, 1] \mid γ(x) = R_V(t(a))\}$$

and

$$δ : [0, 1] \to X,$$

$$δ(x) := γ(\gamma(x_0 + (1 - x_0)x)).$$

Then $δ$ is a continuous path from $R_V(t(a))$ to $R_V(f)$.

We now proceed by theorem 1009 (Bounded transfinite induction) bounded by the number of arcs to define an undirected path $p$ from $s$ to $f$. Since the path $γ$ is simple, it does not intersect itself and hence the image of the arc $a$ at each step will not be in $δ$. □

**Corollary 1349.** A finite quiver is weakly connected if and only if its geometric realization is path connected.

**Proof.** Follows from proposition 1348. □

**Corollary 1350.** A finite undirected multigraph is connected if and only if its geometric realization is a path connected.

**Proof.** Follows from corollary 1349. □
17. References


Brian M. Scott. *Showing that a Transitive Set of Transitive Sets is an Ordinal.* Aug. 12, 2014. URL: https://math.stackexchange.com/a/1057102/229174.


Michael Stoll. *How to prove that two univariate polynomials are always algebraically dependent?* Nov. 12, 2014. URL: https://mathoverflow.net/a/189191.

Todd Trimble. *Is it possible to define a closure operator in terms of partial ordering?* June 11, 2011. URL: https://mathoverflow.net/a/68849.


Ittay Weiss. *Can it happen that the image of a functor is not a category?* June 6, 2013. URL: https://math.stackexchange.com/a/413144/229174.